Situation Awareness in Medical Practice

Dr Helen Higham MBChB, FRCA, SFHEA

Pembroke College

University of Oxford

Michaelmas Term, 2018

Thesis submitted for the degree of Doctor of Philosophy in Clinical Neurosciences

ABSTRACT

Situation Awareness in Medical Practice

Submitted for the degree of Doctor of Philosophy in Clinical Neurosciences, Michaelmas term 2018

Helen Higham, MBChB, FRCA, SFHEA

Pembroke College, University of Oxford

Situation awareness (SA) is a cognitive skill which has been studied extensively in military and industrial settings. Recent studies have shown that errors in SA commonly underpin adverse incidents in healthcare but there is little data to improve our understanding of how to measure or improve SA in clinical settings.

The aim of this thesis was to draw attention to the importance of SA in healthcare and to provide insights into how we might better train multidisciplinary teams in acute care settings. Error in healthcare and research into SA were reviewed and a novel, holistic method for the analysis of the role of SA in the evolution of error in an acute care setting was devised. A systematic review of tools for the measurement of non-technical skills (NTS) was undertaken and this informed a study to assess the reliability and usability of such tools in the measurement of SA. The final study involved the analysis of a variety of different techniques for the measurement of SA in teams from adult intensive care (AICU).

The results revealed that SA errors were present in 96% of serious incidents in a large teaching hospital in the NHS. Challenges were highlighted in the measurement of NTS in healthcare including: 76 measurement tools, revealed by the systematic review, with great variability in quality of design and psychometric testing; low levels of reliability amongst expert raters using these tools and limited evidence of validity for direct and indirect measures of SA in simulated scenarios for teams from AICU.

This work has revealed that SA errors are common in acute care settings and that there are significant challenges in the reliable measurement of SA. Future work should focus on improving measurement of SA and intelligent targeting of teamwork training which highlights the importance of SA and forms part of a system wide safety strategy in the NHS.

ACKNOWLEDGEMENTS

Writing my thesis has been an extraordinary journey both personally and academically. There are so many people without whom none of it would have been possible and who deserve enormous thanks. First to my wonderful family who have been such an amazing support. To my mother and father for reminding me how proud they were and most especially my son Charlie who has had a distracted mother for the past three years.

My utterly fantastic supervisors, Charles and Duncan (my glass half full and half empty respectively), have provided immense wisdom, shown incredible patience and delivered sound advice always at the right moment – I am so lucky that they agreed to do it!

To Liz for being the one to suggest doing this in the first place and for her customary kindness and wisdom when I thought I couldn't , and Wendy (the most organised woman on the planet) who has been an incredible support to me and enabled me to do my job and my thesis.

The SASi study was a huge undertaking and would not have been possible without the amazing team of technicians at OxSTaR (Alan, Charlie and Russ) who never said they couldn't and always found a way. Laura's unbridled enthusiasm ensured we got the work done (and so much more!). The education team in AICU, especially Trevor Venes along with Jonathan Chantler and Stuart McKechnie were so supportive of this work – thank you.

To Irene Tracey for her unfailing support and encouragement (and for the vivas!) and to Mica Endsley for so willingly sharing her knowledge of situation awareness.

To Paul Greig for his inspiration and for listening to me when I needed to off-load and John Rutherford for his generosity and gentle encouragement.

To George Hadjipavlou, Peng Liu and Kilem Gwet for their advice on statistical analysis and Neal Thurley at Bodleian libraries for helping me with literature searches.

Richard Canter was the first to point out that this could be done on top of my day job and has continuously reminded me of all the good things that would come as a result of persevering – huge thanks to you (and Tom Revington, of course).

To my friends and colleagues in the OUHT who have kept me sane and shown polite (sometimes bemused) interest in what I was doing, whilst reminding me of the relevance of the work to my clinical practice. Particular thanks to my dear friend Mhairi and Jason (for the laughter and distractions).

My colleagues in the Patient Safety Academy and ASPiH have all been incredibly supportive of this work and tolerant of my absences. Peter has been an absolute rock, Lauren a wonderful sounding board and Bryn got me through the first publication - thank you all so much.

Clare Dollery has become such a wonderful friend through this whole process. She has felt my frustrations and enjoyed my success and always been there with a glass of wine.

My beloved (retired but you wouldn't know it!) Head of Department and second father, Pierre, has never doubted I could do it and provided such generous and practical support in proof-reading my work, I'm so very glad I came to Oxford for a visit that sunny day in June 1999.

Last, but most definitely not least, to my dear friend and number one supporter Rosie: what on earth would I do without you? I owe you more than I can ever repay.

ABBREVIATIONS

AAGBI	Association of Anaesthetists of Great Britain and Ireland
ABP	Arterial blood pressure
AICU	Adult Intensive Care Unit
ALS	Adult Life Support
ANTS	Anaesthetists' Non-technical Skills
ANTS-AP	Anaesthetic Non-Technical Skills for Anaesthetic Practitioners
ASPiH	Association for Simulated Practice in Healthcare
BARS	Behaviourally Anchored Rating System
CEM	College of Emergency Medicine
CHFG	Clinical Human Factors Group
CRM	Crisis (Crew) Resource Management
Datix	Patient safety software (including an incident reporting system)
DM	Diabetes Mellitus
ECG	Electrocardiogram
ED	Emergency Department
EM	Emergency Medicine
EPR	Electronic Patient Record
ET	Endotracheal
EWTD	European Working Time Directive
GEMS	Generic Error Modelling System
GDTA	Goal Directed Task Analysis

GMC	General Medical Council
GRS	Global Rating Scale
HAT	Hospital Acquired Thrombosis
HRO	High Reliability Organisation
K-BM	Knowledge-based mistake
NAP4	4th National Audit Project (RCoA)
NCEPOD	National Confidential Enquiry into Perioperative Death
NHS	National Health Service
NHSLA	National Health Service Litigation Authority
NIBP	Non-invasive Blood Pressure
NOTSS	Non-Technical Skills for Surgeons
NPSA	National Patient Safety Agency
NRLS	National Reporting and Learning System
NTS	Non-technical skills
OCCG	Oxfordshire Clinical Commissioning Group
OSCAR	Observational Skill Based Assessment tool for Resuscitation
OTAS	Observational Teamwork Assessment for Surgery
OUHT	Oxford University Hospitals NHS Foundation Trust
OxSTaR	Oxford Simulation Teaching and Research
PEA	Pulseless Electrical Activity
PSA	Patient Safety Academy
PU	Pressure ulcer

R-BM	Rule-based mistake
RCA	Root cause analysis
RCoA	Royal College of Anaesthetists
RCS	Royal College of Surgeons
SA	Situation awareness
SAGAT	Situation Awareness Global Assessment Technique
SASi	Situation Awareness in Simulation
SART	Situation Awareness Rating Tool
SBE	Simulation based education
S-BM	Skill-based mistake
SIRI	Serious incident requiring investigation
VTE	Venous thromboembolism
WHO	World Health Organisation

TABLE OF FIGURES

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION – AN OVERVIEW OF ERROR IN HEALTHCARE

1.1 INTRODUCTION: ERROR IN HEALTHCARE

"Patient safety problems exist throughout the NHS as with every other healthcare system in the world."

Don Berwick.[1]

The past three decades have seen a steady rise in the study and understanding of error in healthcare and of the strategies that could be used to improve safety. Safe care of complex or acutely unwell patients requires high levels of proficiency in both technical and non-technical skills (NTS), such as situation awareness, communication and teamwork. Human beings working in complex, dynamic environments make mistakes and healthcare professionals are no exception. In the past ten years, alongside my clinical practice in anaesthesia, I have taught over 4,000 people about the nature of error in healthcare and how to prevent it. In the safe environment of the simulation centre many of the participants have willingly shared their personal experience of mistakes in clinical practice and have highlighted the difficulties they have faced in coming to terms with error and the challenges of reporting and learning from it.

This chapter will introduce and explore error in healthcare including a brief history of the research into its causes. The human behavioural aspects of error will be considered with a particular focus on NTS. The chapter will conclude with a consideration of the impact of error through the description of a case history in an acute care setting and elucidation of the key points at which errors occurred and why.

1.2 EVIDENCE OF ERROR IN HEALTHCARE

Evidence confirming the existence of error in healthcare has been available in anaesthesia since the 1950's[2–7]and more recently in the wider healthcare community.[8–11] Furthermore, deaths secondary to avoidable error in healthcare have been estimated at 195,000 per annum, the third biggest killer in the United States after heart disease and cancer[12] and similar estimates are described in many of the developed healthcare systems in the world.[13–15] Awareness of the prevalence of error in anaesthetic practice, and the courage to reflect constructively on the underlying causes, has led to pioneering research and training to improve safety. Methodologies to study risk and reduce error have been developed far in advance of the other medical specialities and before publication of documents such as "To Err is Human" by the Institute of Medicine in 1999[16] and "an Organisation with a Memory" in the UK in 2001[17] which led to greater awareness of the issue of avoidable harm amongst the wider healthcare community.

Of course, the most obvious and important impact of error in healthcare is on the patients and their families who suffer the physical and psychological impact of the adverse event. In the NHS the tragic death of Elaine Bromiley in 2005, during anaesthesia for elective ENT surgery, brought the inadequacies of human factors training in healthcare into sharp relief.[18] Elaine's husband, Martin Bromiley, is an airline pilot and his subsequent focus on enabling learning from the errors that were made in his wife's case led to the foundation of the Clinical Human Factors Group (www.chfg.org.uk) and a significant and sustained increase in the awareness of the importance of human factors in the NHS.[19]

The healthcare workplace has changed rapidly and dramatically over the past few decades. Advancements in the technologies and devices used to support patient care; the increasingly

complex comorbidities in our patients; the move to electronic systems of record keeping and changes in patterns of working coupled with increasing pressure to achieve targets, make healthcare a more challenging environment in which to work than ever before. It is in these complex and dynamic settings with multiple interdependencies and the requirement for rapid decision making in uncertain situations that healthcare professionals must strive to prevent error.

1.3 LEARNING FROM HIGH RELIABILITY ORGANISATIONS

There are numerous other industries where challenging work conditions are also regarded as potentially hazardous or high risk. Professor Karlene Roberts, who has studied the design and management of organisations in which errors can have catastrophic consequences, states: "to identify these organisations one must ask the following question, 'how often could this organisation have failed with dramatic consequences?' If the answer to the question is many thousands of times the organisation is highly reliable."[20] These so called High Reliability Organisations (HROs) offer useful lessons to be drawn from their considerable efforts to drive down accident rates through research into human factors and implementation of novel systems design and monitoring strategies. Successful examples of HROs include the nuclear fuel industry, chemical industries and civil aviation (e.g. learning from the egregious events at Chernobyl, Bhopal and Tenerife airport respectively).[21] James Reason and Charles Perrow provide eloquent insights into how such events are caused by a concatenation of behavioural and systems failures leading ultimately to catastrophe.[21–23] Rene Amalberti has studied "ultrasafe" industries and compared them with what he calls "unregulated and dangerous enterprises".[24] Regulated systems include driving and chartered flights whereas dangerous systems include mountain climbing and bungee jumping (which are associated with accident

rates of 1 per 1000 events). Figure 1-1 from Ferner's paper on medication error[25] extrapolates from work by Amalberti[24] and Leape[26] to compare the safety performance in healthcare with other high risk enterprises. Unfortunately, healthcare is currently more closely aligned with mountain climbing than civil aviation.

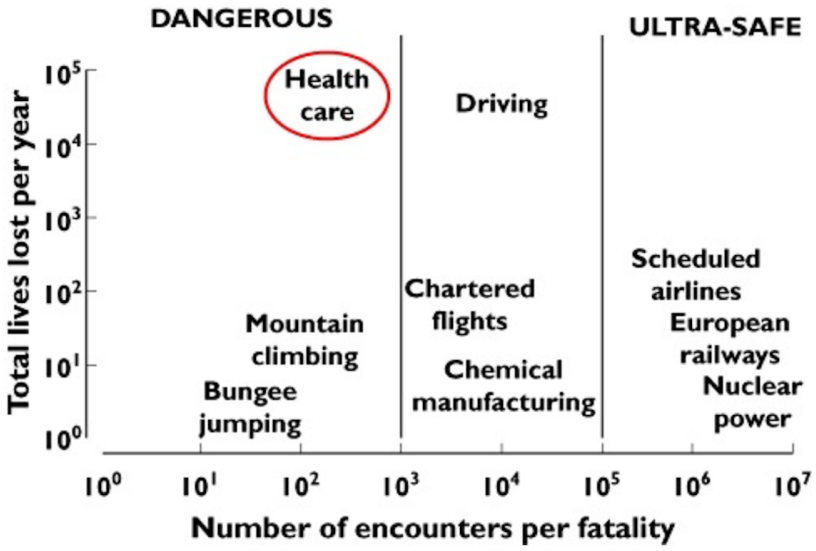


Figure 1-1:The total number of lives lost per year plotted against the number of encounters per fatality on a log-log scale. (reproduced from 'An agenda for UK pharmacology: medication errors' (Br J Pharmacol 2012; 73(6):p912 with permission of Wiley publishers)

One of the clear success stories in healthcare is found in anaesthetic practice in which there has been considerable improvement in patient safety and mortality over the past 50 years. Early reports on patient mortality suggested that deaths directly caused by anaesthesia were in the region of 1-2 per 10,000,[2,3] that figure is now closer to 1 per 100,000.[27] The reasons for this improvement are multifactorial and related both to scientific developments in the field as well as a broader and stronger focus on clinical governance and standards of practice. In more recent years other specialities where good teamwork in time-pressured situations is vital, such as emergency departments and intensive care units, have focused on strategies to improve safety including training to develop personal NTS alongside improving team capabilities in the management of critical situations.[28,29]

The lessons learned from the study of error in HROs and the successful implementation of systems improvement and human factors training are now starting to be adopted in healthcare and there is emerging evidence of performance and outcome improvement.[30,31]

1.4 HUMAN FACTORS

Human factors expertise has been successfully employed in many industries outside healthcare to improve safety (e.g. for the investigation of error, the delivery of effective systems interventions and the design of training programmes for staff). Human factors has been defined by the Health and Safety Executive as the "environmental, organisational and job factors, and human and individual characteristics, which influence behaviour at work in a way which can affect health and safety".[32] Whilst originally intended to address the well-being of the worker, the impact of a Human Factors approach to systems design is readily extended to patient safety, productivity and efficiency in the healthcare context. The two broad domains of study under this umbrella are human behaviour and systems analysis (with considerable interdependency between the two). They provide significant opportunities for simulation-based approaches to help support the integration of human factors into education and professional practice.

The discipline and study of human factors began in earnest during the Second World War when it was recognized that as scientific advancements in warfare delivered more technologically complex equipment for humans to use the expected benefits were not realized when soldiers were unable to understand or use the more complicated weaponry. An obvious analogy exists in medicine where equipment, medication and treatment modalities change rapidly without the clear focus that exists in other domains to design systems around the human-device interface. An increasing desire for providing information technologies to reduce reliance on

traditional paper based documentation matches similar trends in society but healthcare does not yet have the embedded methodologies or culture of pursuing a user-centred approach to the design of human/device interfaces. Providing optimal care for patients presenting with complex problems and extensive comorbidities in overburdened under-resourced and inadequately designed systems leads to error, potential harm and waste in the use of resources to support care.[21]

1.5 THE EVOLUTION OF ERROR

Errors leading to harm in healthcare are almost always the result of a combination of behavioural and systems problems, for example: fatigue or stress in staff combined with standard operating procedures which are not easily accessible or fit for purpose in a system where resources are limited and work environments outdated provide ideal circumstances for mistakes to occur. Reason describes the "human-as-hero" in these conditions[33] i.e. it is far more often the case that a healthcare professional will notice a discrepancy and prevent an error than miss one and harm a patient. However, errors do happen and when they harm patients there is a moral and professional imperative to understand what went wrong and learn from it. Reason's "Swiss cheese" model (Figure 1-2) of the evolution of error has been widely applied in industry[21] and more recently healthcare[34] and describes how a series of latent conditions which stem from the higher echelons of organisations (e.g. cultural and systems problems such as the design and construction of the work environment and the allocation of resources) combined with active failures which occur more proximally at the human-system interface (e.g. a suboptimal working climate leading to poor teamwork or inadequate communication) can allow the holes in the Swiss cheese to line up and let an error through. Whilst the Swiss cheese model has gained widespread acceptance, for the purposes of considering ways of error prevention

the author prefers to use a “chain of error” because it is conceptually more satisfying to think of breaking a chain and preventing an error. The example included at the end of the chapter refers to Reason’s concepts of active and latent failures but in the form of a chain of events leading to the error.

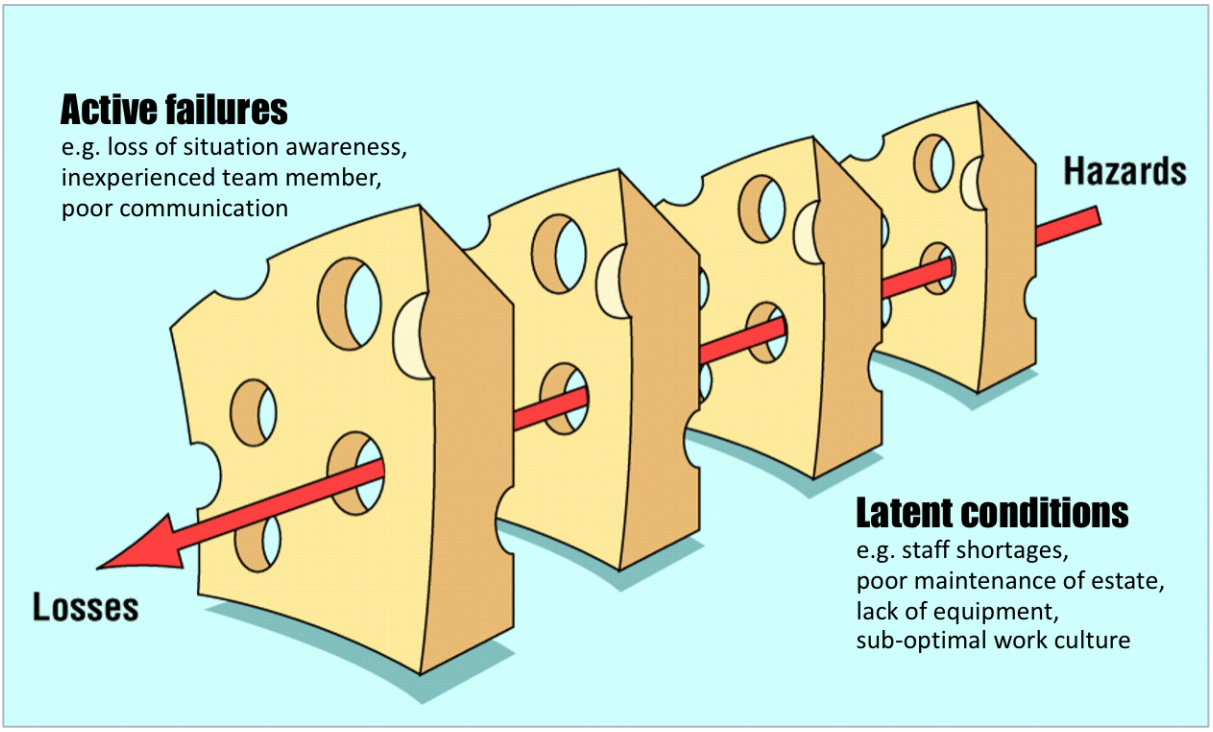


Figure 1-2: Reason's Swiss Cheese Model showing how error occurs when a combination of active failures and latent conditions align (adapted from Reason, BMJ 2000).

The study of human error has a long history in psychology[21] and the associated cognitive constructs which underpin error include perception, attention, long-term and working memory, and the use of heuristics as cognitive short cuts. These key domains will now be discussed in more detail.

1.5.1 PERCEPTION, ATTENTION AND ERROR

Perception and attention are closely related but are not the same thing: perception is a subjective conscious awareness of a stimulus (e.g. the sound of a monitor alarm) whereas attention requires direction of sensory or cognitive resource towards that stimulus (recognising

the alarm and turning towards the monitor to determine the cause). Although sub-optimal vigilance and failures of perception in clinical situations are known to cause error,[35] few healthcare professionals are aware of the problem.

It would clearly be impossible to perceive in detail all the myriad facets of the world we exist in – the brain does not possess the necessary processing power and must filter unnecessary information and focus on meaningful data.[36,37] Furthermore, human attentional resources are limited and may be disrupted by the challenges of multimodal sensory inputs and the requirement to monitor several inputs at once[38,39] – factors which commonly coexist in medical emergencies. As mental workload increases attentional focus narrows (even for experienced operators)[40] and increasing age[41] and mood[42] also impact perception. Flaws in our ability to detect alterations in our environment are well recognised[43] and failures of perception may also result from lapses in conscious awareness (e.g. whilst engaging in repetitive tasks or during long periods of monitoring)[44]. Failures of perception and attention are common in what Reason has termed Skills Based errors (slips and lapses).[21]

1.5.2 MEMORY AND ERROR

Long-term memory is "a vast store of knowledge and a record of prior events, and it exists according to all theoretical views",[45] these memories are neither entirely accurate nor absolutely complete but are a vital part of higher order cognitive skills such as situation awareness[40] and decision making.[46] Short term memory holds information that is easy to access but limited both by amount[47] and duration[48] of content availability. The construct and mechanisms of working memory have been the subject of considerable debate but Baddeley has proposed that "a dedicated system maintains and stores information in the short term" and that this system underlies and serves human thought processes.[49] Cowan argues that short

term memory is a part of working memory[45] and it will be considered as such for the purposes of this thesis. There are limits to how much working memory can cope with[47,50] and these limitations vary between individuals[51] and with age,[52] stress[53] and experience.[50,54]

Two abstract constructs associated with long-term and working memory are vital in understanding and interpreting the world around us: schemata and mental models. There has been some ambiguity in the literature regarding the meaning of the two terms[55–58] and how they are distinguished.

Reason has crystallised the extensive body of work on schemata and explained them as, "higher order, generic cognitive structures that underlie all aspects of human knowledge and skill. Although their processing lies beyond the reach of awareness, their products – words, images, feelings and actions - are available to consciousness. Their encoding and representational functions include lending structure to perceptual experience and determining what information will be encoded into or retrieved from memory. Their inferential and interpretive functions go beyond the given information, allowing us to supply missing data within sensory or recalled information."[59]

Schemata may be constructed from direct or indirect experience (a doctor either experiences a particular clinical situation himself or hears about it from a colleague) and training. They are stored in long term memory ready for activation at a future point and are vital as reference points for the construct of mental models.

Mental models are internal representations of real, hypothetical or imaginary situations. They are created from perceptual input and reliant on existing schemata to help make sense of a current situation.[57,60,61] Brewer[61] explained the difference between schemata and mental

models as fundamentally temporal in nature i.e. mental models are created within working memory in the moment whereas schemata exist in long-term memory. This is how they will be defined for the purposes of this thesis.

1.5.3 HEURISTICS AND ERROR

Heuristics are the simple rules of thumb used on a daily basis to assist us in coming to decisions or making choices. Reason has described humans as "furious pattern matchers"[21] and it has been shown that "humans, if given a choice, would prefer to act as context-specific pattern recognisers rather than attempting to calculate or optimise."[62] Our desire for an easy solution to a problem leads to a tendency to choose the most frequently occurring or recent events and apply rules which worked in those situations. Examples include representativeness (similarity) and availability biases and they have been shown to lack robust reasoning based in fact.[63] One example given in Tversky's and Kahneman's work[63] is as follows:

Consider the following description of a man:

"Steve is very shy and withdrawn, invariably helpful, but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail."

When asked to rate the likelihood of Steve's occupation being one of the following: farmer; salesman; airline pilot; librarian or physician the likelihood of "librarian" being chosen is high. The reason for this is that Steve's characteristics are matched with those of a "typical" librarian (he is *representative* of the type) rather than considering that it is far more likely that he is a farmer because there are many more farmers in the population than librarians.

Our preference for reliance on such heuristics rather than hard facts leads to errors of judgement and poor decisions. Problem solving relies on appropriate and accurate use of

knowledge and stored rules and problem solving errors are common in rules- and knowledge-based mistakes as described below.

1.6 THE GENERIC ERROR MODELLING SYSTEM

Reason has developed a Generic Error Modelling System (GEMS),[21] based on the three performance levels described by Jens Rasmussen,[64,65] to categorise the types of errors humans make. There are three error levels:

- Skill-based (SB) slips or lapses – the result of memory or attentional deficit during actions or activities that are taking place without conscious control (e.g. a distraction during a task such as writing a prescription)
- Rule-based (RB) mistakes - humans prefer to find a solution which matches a similar previously experienced problem: the "I've seen this before" feeling (e.g. low blood pressure = bleeding)
- Knowledge-based (KB) mistakes – the most cognitively effortful level of error where the "problem space" is not easily recognisable (e.g. a set of symptoms or signs representing a disease process never previously seen by the doctor), requires diagnosis of a problem and formulation of an effective solution, often implemented after recurrent failures at the rules-based level are recognised

Table 1-1 summarises the differences between SB, RB and KB errors

Dimension	Skill-based errors	Rule-based errors	Knowledge-based errors
Type of activity	Routine, everyday procedures	Problem solving activities	
Focus of attention	Directed at something other than the task in hand	Directed at problem related issues	
Control mode (automatic/ unconscious or conscious)	Mainly automatic / unconscious Schemata	Mainly automatic / unconscious Stored rules	Limited, conscious processes
Predictability of error type	Largely predictable "strong but wrong" actions	Largely predictable "strong but wrong" rules	Variable (KB errors are more likely in "uncharted territory")
Ratio of error to opportunity for error	Absolute numbers may be high (because SB and RB processes are more prevalent) but are a small proportion of the opportunities for error		Absolute numbers small but opportunity ratio high
Influence of situational factors	Low to moderate: intrinsic factors (cognitive biases, attentional limitations) likely to exert the dominant influence		Extrinsic factors likely to dominate (structural characteristics of task and context)
Ease of detection	Detection usually rapid and effective	Difficult and often only achieved through external intervention	

Table 1-1:Summary of distinctions between skill-, rule- and knowledge-based errors (adapted from Reason[21])

Reason also highlights the importance of change as a feature common to error provoking situations. At the SB level the variants in a situation are usually predictable and recognisable:

- An anaesthetist has just transferred an anaesthetised patient from the anaesthetic room into theatre when the surgeon requests that an additional antibiotic be given (a fairly routine occurrence) the anaesthetist turns to go back into the anaesthetic room and prepare the antibiotic and, on the way, the anaesthetic nurse mentions that the patient had complained of feeling sick after a previous anaesthetic. The anaesthetist forgets the antibiotic whilst preparing an anti-emetic for the patient.

In RB mistakes the changes are predictable but only in the sense that they may have been seen or experienced before - the timing is unknown and the exact form of the problem unspecified. Usually a strategy for mitigation may be available but a good rule misapplied or application of a bad rule may lead to error:

- A swab is inadvertently left in a wound after surgery because the standard operating procedure for counting swabs is not followed properly (misapplication of a good rule)
- A patient is transferred from one site to another with inadequate medical assistance and monitoring (application of a bad rule: the standard operating procedure for the safe transfer of patients is poorly designed and difficult to understand, the patient is inappropriately deemed fit for low dependency transport)

In KB mistakes the changes encountered are not recognisable or planned for and rely on the cognitively effortful and error prone processes of reasoning:

- A patient deteriorates rapidly after extubation on intensive care and the endotracheal tube cannot be repositioned in the usual way (via the mouth or nose). The team involved has not faced a situation this extreme before and the opportunity to site a surgical airway (tracheostomy) at an early stage is missed.

These three fundamental error types are pervasive across all forms of human activity and the perceptual, attentional and cognitive attributes described above are integral to the NTS I will now consider.

1.7 NON-TECHNICAL SKILLS AND ERROR

Rhona Flin (an industrial psychologist at the University of Aberdeen) has defined NTS as "the cognitive, social, and personal resource skills that complement technical skills, and contribute to safe and efficient task performance. They are not new or mysterious skills but they are essentially what the best practitioners do in order to achieve a consistently high performance and what the rest of us do on a good day."[46] She described a further division of these skills into two subgroups: cognitive skills (e.g. decision making, planning and situation awareness) and social skills (e.g. communication and team working). It is worth highlighting that concern exists

around the use of the term NTS[66] to describe such important aspects of professional clinical practice, however, whilst there is currently no universally agreed substitute[67] the term NTS will be used for this thesis.

These skills are mainly abstract in nature and, therefore, can be challenging to recognise and assess. Behavioural markers are observable individual or team behaviours associated with NTS which are usually structured into a set of categories. They are derived from observations, task analysis, interviews and focus groups etc. from the context in which they are used (e.g. theatre staff may be observed and questioned about the behaviours which contribute to successful [or unsuccessful] performance of tasks in the operating theatre). Flin describes five categories[68] along with two that focus on the management of stress and fatigue:

- Situation awareness (SA)
- Decision making
- Communication
- Teamwork
- Leadership

In light of the challenges associated with the assessment of NTS, recommendations have been made for the use of behavioural markers;[69];the training requirements for NTS assessment[70] and for NTS competency based teaching in healthcare.[71]

Error in healthcare is far more likely to be caused by human error than by machine failure and NTS are more commonly implicated than technical skills. It has been estimated that 70-80% of errors in healthcare are attributable to failings in these domains[46]. Situation awareness is vital for making good decisions[40] and errors in SA have been shown to have a significant impact in serious incidents in air traffic control,[72] civil aviation,[73] nuclear power[74] and healthcare.[75–79] Simply put, SA is knowing what is going on around you and the next chapter will explain this

cognitive skill in depth, however, the following case study will provide evidence of its vital role in the evolution of error in healthcare.

This case study is an anonymised summary of an incident investigated by the author. It provides examples of the systems and behavioural errors (including NTS errors) which are usually evident in critical incidents in healthcare.

1.8 PUTTING IT ALL TOGETHER: ILLUSTRATION OF AN ERROR CHAIN IN A CASE STUDY FROM AN ACUTE CARE SETTING

A 68 year old man with a complicated past medical history of diabetes, ischaemic heart disease, heart failure, poorly controlled hypertension and a previous stroke arrived at the Emergency Department in a large teaching hospital with a history of recently worsening shortness of breath. He had only recently moved to the area and his medical notes describing the extensive investigations and treatment he had received for his heart failure were not available. Furthermore, the patient and his family gave conflicting views of his exercise tolerance and independence.

His condition deteriorated to the point that he required prompt endotracheal (ET) intubation and ventilation to support his breathing and he was admitted to a side room on the Adult Intensive Care Unit (AICU). It was difficult to pass the ET tube because the patient was obese and his upper airway anatomy was such that it was very hard to see the opening of his trachea (wind-pipe). Over the next few days on AICU he had a number of investigations including an echocardiogram which displayed very poor ventricular function (severe heart failure) and the consultant who performed the echo made a note that the patient's prognosis was very poor. This message was not passed on accurately to other team members.

The patient's respiratory function gradually improved and he was extubated on day four of his stay on AICU but the potential for deterioration was not anticipated. Unfortunately, the patient's condition worsened rapidly after the extubation and he required emergency reintubation. The staff involved in the incident did not notice the bespoke difficult airway plan placed at the head of the bed and the cramped space in the side room made it impossible to get all the necessary equipment into the room. The reintubation was difficult and the team found it impossible to pass a tube orally (via the mouth) and in this high pressure setting did not recognise the need to progress to a surgical airway using front of neck access (FONA). At this time there was no regular team based training in the management of emergency airway situations. During the attempt at reintubation the patient desaturated and suffered a cardiac arrest. Two additional senior doctors arrived and highlighted the need to proceed to FONA. This was a very challenging procedure to undertake in an extremely high pressure situation. The patient recovered from the cardiac arrest but had suffered a period of hypoxia and his heart failure worsened over the following days leading to his death four days later.

The investigation found that there were a number of contributory factors in the evolution of this incident. These factors and the relevant error type according to GEMS are shown in Figure 1-3 and explained below.

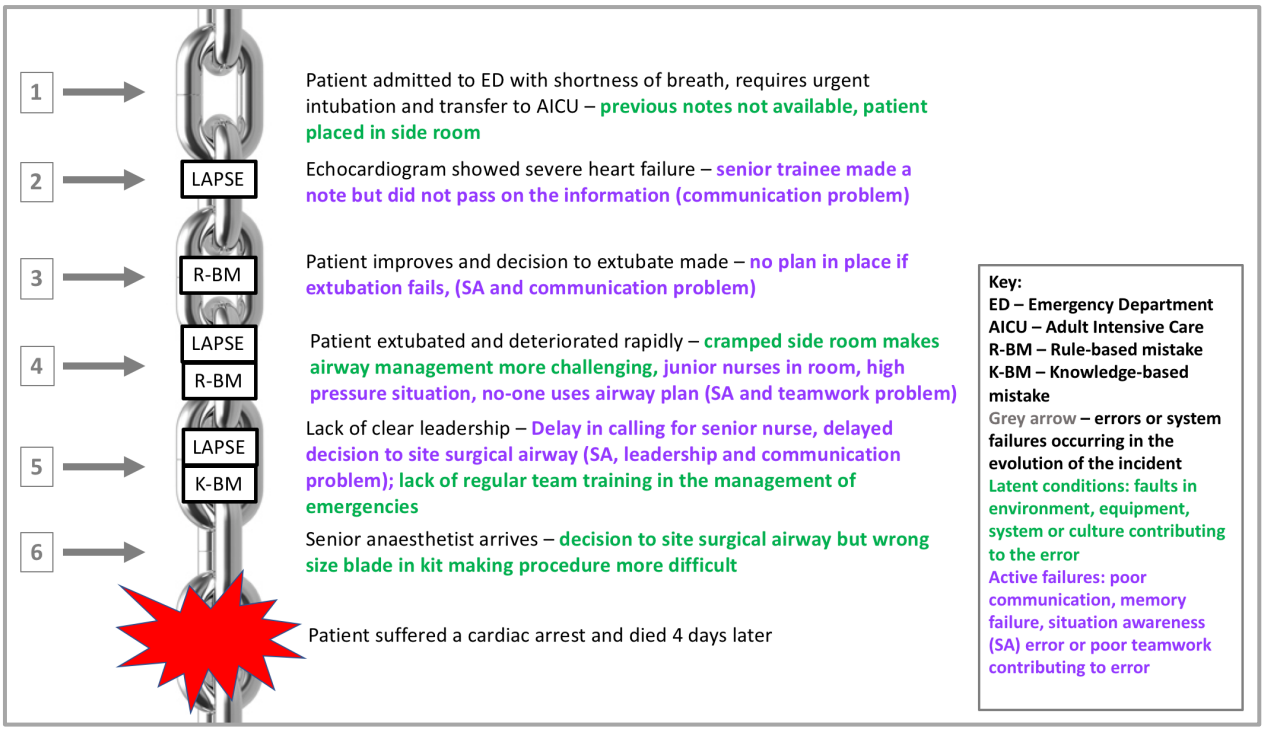


Figure 1-3: Error chain depicting latent conditions, active failures, non-technical skills and GEMS designation for a SIRI involving a critically unwell patient on AICU.

1. Latent conditions:

- Lack of continuity of note keeping across the NHS – this patient's records were in another hospital and no transfer of information had occurred
- The patient was placed in a side room on AICU as this was the only available space at the time. The intensive care unit is sited in a part of the hospital which was designed and constructed in the late 1970's and space requirements for intensive care are now larger.

2. Active failure – skill-based lapse:

- The consultant performed an echocardiogram and recorded the result on the electronic system and told the senior trainee. The trainee made a note but did not pass the message directly to the consultant in charge of AICU on that day

3. Active failure – rule-based mistake

- Failure of the team to follow rules on safe extubation caused by lack of anticipation of problems and no explicit team communication about the plan if reintubation was needed (SA error – team members had different "mental models" of the situation and did not share them, see Chapter 2).

4. Latent condition and active failure - rule-based mistake:

- small side-room made management of the difficult airway more challenging
- In the high pressure setting teamwork and SA was sub-optimal and good rules were forgotten (use of the airway plan)

5. Latent condition and active failure – knowledge-based mistake:

- There was no regular programme of simulated emergency training to support the development of teamworking skills in a crisis and the stress of the moment, coupled with the lack of experience in siting a surgical airway, led to a loss of SA and impaired decision making
- Leadership during this very stressful incident leadership was unclear and decisions to call for senior nursing assistance and to site a surgical airway were delayed

6. Latent condition:

- When the decision to use FONA was made it was discovered that a small scalpel (size 15 rather than size 20) had been placed in the surgical pack. The surgical packs for FONA were new to the Trust and had never been used before on AICU.

This incident is an example of a succession of factors (comprising latent conditions and active failures) which led to the unfortunate death of a patient on AICU. It is one of the most egregious circumstances which can occur in an intensive care setting and provides an example of the type of serious incidents this thesis will examine closely.

1.9 CONCLUSION

This chapter has given an overview of error in healthcare and the human factors approach to reducing error taken by HROs. The human behavioural reasons underlying error have been explored and the GEMS model for error categorisation has been explained and used to defined errors in healthcare. These error types have been linked to NTS errors and an investigation of a serious incident on AICU revealed that SA errors occurred on multiple occasions in different team members. The next chapter will explore SA in more detail and consider SA errors and their role in critical incidents.

CHAPTER 2 DEFINITION AND OVERVIEW OF SITUATION AWARENESS

2.1 DEFINITION OF SITUATION AWARENESS

Situation awareness is an abstract construct describing a cognitive process in individuals and is relatively new phenomenon in human factors research. Endsley has defined SA as an *individual's* "perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future", more informally put it is "knowing what is going on around you."[80] As a concept, it has been in existence since the First World War[81] but it was first studied in depth in aviation research six decades later[80,82,83]. There has been extensive debate about its definition and relevance and around the cognitive pathways underpinning the development of SA. [84–87] Pew highlighted the importance of being able to separate SA meaningfully from other aspects of performance and distinguish it from concepts such as perception, workload, and attention.[40] Others have focused on understanding SA as a construct in changing environments;[88–90] the importance of distinguishing process from product[91,92] and developing a mathematical model to describe SA.[93] More recently it has been divided into individual, team and system SA.[94] Human factors experts in planning, training and research from aviation, road transport, the military, accident investigation and nuclear power have subsequently concluded that it is a vital and measurable component of successful performance in humans[95] and, consequently, there has been an upsurge of interest in healthcare settings.[96–100]

2.2 UNDERSTANDING SA

Historically the maintenance of good SA has been important for the survival of the human race but, whilst awareness of imminent attack is not a concern in modern healthcare, doctors and nurses exist and work in far more complicated systems than ever before. Rapid developments in information technology have profoundly altered the interfaces through which we perceive and access information to support SA; work environments are more complex and dynamic and the number of interdependencies between teams in the provision of complex care pathways (e.g. major trauma units) has increased enormously.

Psychologists and engineers in military and civil aviation have a particular interest in human performance under pressure and have spearheaded the research into situation awareness.[101–103] Endsley's definition and three level model of SA (see below) is by far the most cited and used.[80,95,104]

All healthcare professionals working in acute care settings must be able to respond to rapid deteriorations in the condition of their patients - to do so requires the gathering of data from the patient, from colleagues, from the environment and from an increasingly complex array of medical monitors and devices (with no standardised layout of data). Endsley has referred to the challenge of acquiring and utilising the relevant data from this increasingly data rich environment as the "information gap" (Figure 2-1).[80] It is not the case that more data leads to more information and better decisions - overwhelming quantities of data may overload cognitive processing power and lead to confusion and incorrect actions based on flawed SA.

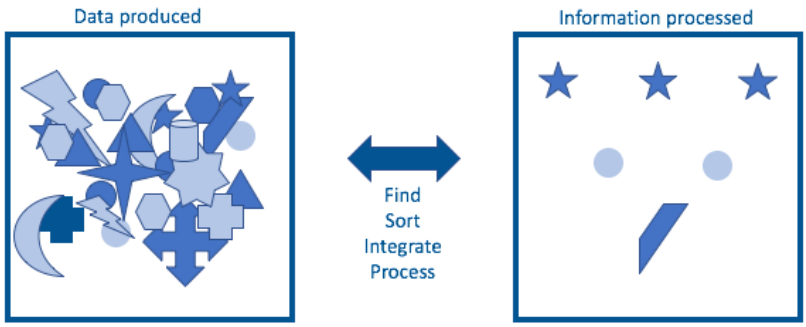


Figure 2-1: Difference between data produced and information processed during situation assessment (the "information gap"). Adapted from Endsley[105]

SA is not a static construct and must be regularly updated to maintain accuracy and support appropriate decision making. The *process* of acquiring SA has been termed *situation assessment* and the *product* of that assessment is *situation awareness*.

Figure 2-2 describes the various processes (situation assessment) involved in an individual's development of situation awareness in dynamic settings and the factors which may affect its accuracy. Endsley divided SA into three levels all of which may be influenced by internal (e.g. memory, stress or fatigue) or external (e.g. system or environmental) factors and by goals and expectations.[40] Situation assessment will also vary according to professional background and the context of the situation. This model is not designed to mimic neurophysiological connections or the complexity or speed of the cognitive processes which underpin the development of SA but it is helpful in considering how SA may deteriorate and how systems or training interventions may be targeted to improve SA.

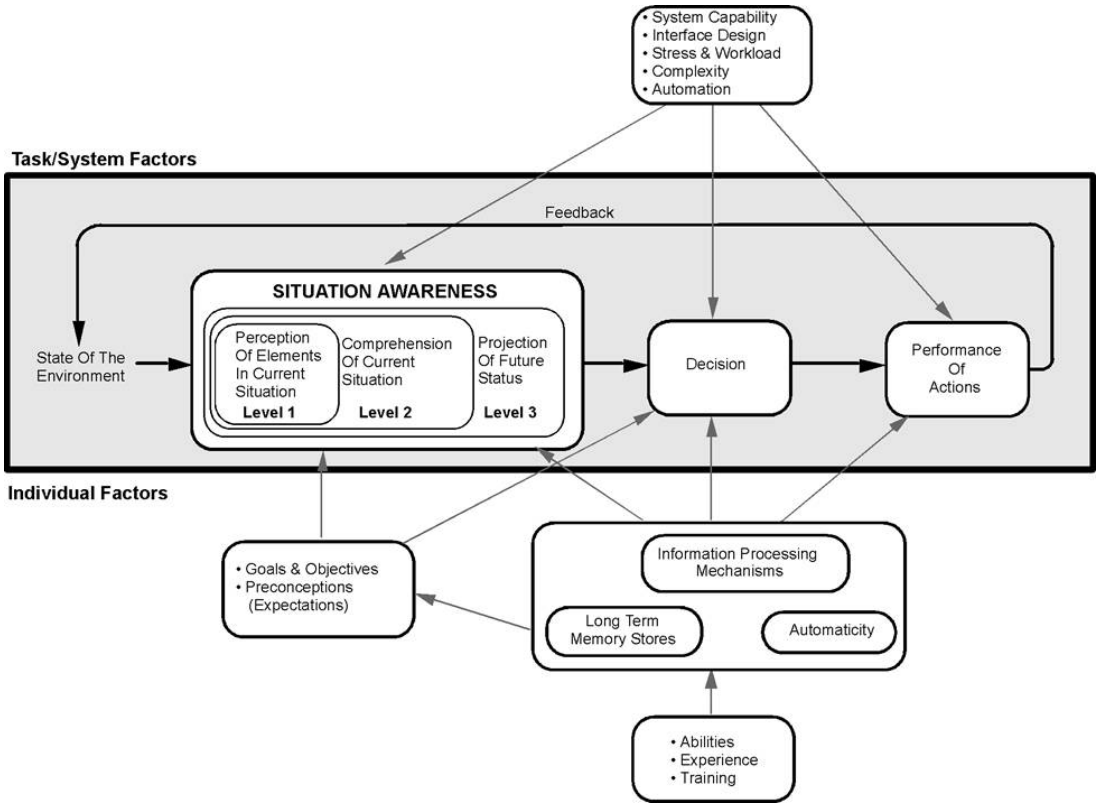


Figure 2-2: Model of SA in dynamic decision making from "Measurement of Situation Awareness in Dynamic Systems" by M.R. Endsley. In Human Factors, 1995; 37(1): 32-64. (Permission not necessary for reuse in a thesis).

2.2.1 LEVEL 1 SA: PERCEPTION

Level 1 SA describes the first step in the process of developing SA. In healthcare data are gathered, using all five sensory modalities, from our surroundings, the people (including the patient, members of the patient's family and co-workers), and monitoring equipment.

Table 2-1 gives examples of the elements of SA at all three levels which may be required in the management of an in-hospital cardiac arrest (a challenging but frequent occurrence in hospital practice requiring involvement of a multidisciplinary team some of whom [e.g. porters] may not have a clinical background). A doctor gathers information in a cardiac arrest by looking for a visible physical obstruction in the airway such as regurgitated stomach contents or listening for air entry in the lungs. Team members may provide additional information on the patient's condition and family members may provide more detail on the patient's long-term health and

quality of life. Monitoring equipment will provide data on the patient's heart rate, blood pressure and level of oxygen saturation and blood tests on respiratory or renal function etc.

Level 1 (perception)	Level 2 (comprehension)	Level 3 (projection)
Information from patient / monitors: Does the patient have a patent airway? Is the patient being ventilated? What is the cardiac rhythm?	**Comprehension of the situation may include:** The patient requires intubation to ensure adequate ventilation (this may require additional skills). Diagnosis of the type of heart rhythm the patient is in.	**Projection of requirements for treatment:** An anaesthetist or other specialist personnel will be required, this must be arranged quickly.
Information from the environment: Is the patient accessible for treatment (i.e. on a bed/trolley)? Is there adequate resuscitation equipment in the vicinity? Are staff competent to assist with resuscitation?	Understanding if the patient is in a suitable place to provide resuscitation (e.g. may require moving from a chair to a bed). Understanding if the necessary resuscitation equipment is available (e.g. a defibrillator If the rhythm is shockable). Deciding if the staff present have the necessary skills to provide care (is a call for help needed?).	Defibrillation must be prioritised and delivered early – a member of staff should be sent to get the necessary equipment promptly.
Information from staff and patient record etc.: What is the most up to date information on the condition of the patient (usually available from the responsible nurse or doctor)? Is there any additional information available from notes/results/investigations? Is there any additional information available from family members or carers? Knowledge of the patient's preference for treatment in the event of a cardiac arrest (some may not wish to be resuscitated).	Knowledge of the patient's co-morbidities or recent interventions that may impact treatment (e.g. if they have severe heart failure resuscitation is unlikely to be successful or if they have had surgery there may be an unacceptable risk of bleeding if anticoagulants are necessary for a heart attack). Additional information from the family about the patient's medical history (if notes are unavailable vital information e.g. accurate assessment of the exercise tolerance of a patient).	Prediction of the requirements for after care will require advance planning (e.g. it may be necessary to move the patient to a catheterisation lab for coronary stents or to an intensive care unit for invasive monitoring). Prediction of likely treatment requirements will need to be balanced against the existence of comorbidities which require additional management or recent surgery which would add increased risk to some procedures which might be required after cardiac arrest.

Table 2-1: Example elements at three levels of SA for a doctor treating a patient who has suffered an in-hospital cardiac arrest

The gathering and storing of information requires an individual to direct attention at key pieces of information and then store them in working memory but attentional resource and working memory are limited and also vary between individuals (see Chapter 1). Emergency care settings, such as the case described at the end of Chapter 1, present scenes with theoretically unlimited amounts of visual and auditory data for the clinician to sift through. Cognitive processing power would be rapidly overwhelmed if the brain did not divert attention to key areas of interest relevant to the goals of the task in hand. An individual's attention will be directed more frequently to important or rapidly changing parameters (e.g. heart rhythm or blood pressure in a cardiac arrest) and less frequently to items which don't require revisiting often (e.g. the amount of oxygen in the cylinder being used) but filtering of information in situations such as these can lead to loss of relevant data.[43] Studies have also revealed that as cognitive workload increases, attentional focus narrows[106] i.e. we are more likely to miss things that matter if we are trying to do too many things at once, particularly in high pressure settings.[105]

Processing of information is additionally affected if the same modality, or response mechanism is required:[107]

- Multiple visual stimuli (e.g. two traces on a monitor representing oxygen and carbon dioxide levels and a paper 12-lead ECG recording) occurring simultaneously are harder to process than a visual, auditory and tactile stimulus occurring simultaneously (e.g. one trace on a monitor, a verbal report of current blood pressure from a nurse and a hand on the patient to feel their temperature).
- When the response to information needs to be delivered the same way e.g. a change in heart rhythm needs to be expressed verbally to the team and, at the same time, a

response to a report from a team member requires a verbal command.

Examples of how inaccurate SA at level 1 may develop as a result of internal or external factors include:

- Data are misperceived e.g. mishearing data in a dynamic, noisy situation
- Working memory has limits and overload leads to loss of what may be vital information
- Data are not monitored or observed due to inadequate search strategies
- Poorly designed work environments lead to difficulty detecting relevant data (e.g. small side rooms, badly lit wards)
- Poorly designed monitoring equipment leads to difficulty detecting relevant data e.g. too much data on one screen or too many screens makes finding relevant information difficult
- Information stores are inaccessible e.g. data are not present because patient notes have been lost or electronic patient systems are not accessible
- Staff have not been trained in the use of search strategies

Research in aviation, driving and healthcare settings has shown that most errors in SA occur at level 1.[78,79,108–110] This will be discussed in more detail below.

2.2.2 LEVEL 2 SA: COMPREHENSION

Level 2 SA describes the development of a meaningful construct of what is happening in a given moment by the integration of disjointed elements collected as described above in the context of individual and team goals and expectations (see below). It requires the individual to prioritise and order information (analogous to being able to comprehend meaning in words as opposed to just reading them) i.e. the doctor in a cardiac arrest would see a line on a monitor

representing a heart rhythm and would understand the type of rhythm disturbance and the underlying causes.

Memory again plays a vital role in the development of level 2 SA and stored schemata serve as templates for the development of mental models “in the moment” (see Chapter 1).

Creating and testing mental models requires effort particularly in rapidly changing situations such as the cardiac arrest described in Table 2-1 or the incident highlighted at the end of Chapter 1. Schemata can provide “shortcuts” to decrease the cognitive workload particularly for experienced personnel. However, in the example given in Table 2-1 the first doctor to arrive is often quite junior and may have little experience in the management of a cardiac arrest.

Inaccurate SA at level 2 may, therefore result because:

- Inaccuracies are present at level 1 e.g. too much information is presented in a non-standardised format (either on electronic screens or on paper) which makes it more difficult to find salient information to fit together in a recognisable pattern
- Stress or fatigue impact working memory and other cognitive processes reducing the ability to build an accurate mental model of the situation
- Lack of adequate (experiential) training or clinical experience reduces exposure to problem solving (diagnosis) under pressure and makes it less likely a clinician will have schemata to apply to a particular crisis

The error to opportunity ratio is, therefore, higher at the more cognitively effortful level 2 and level 3 SA than at level 1 (see Chapter 1).

2.2.3 LEVEL 3 SA: PROJECTION

Level 3 SA requires extrapolation from the construct developed at level 2 into what might happen next and what actions may need to be taken as a result. The synthesis of a plan of action that will result in a positive resolution of the situation requires accurate level 1 and level 2 SA. Inaccurate SA at level 3 may result because:

- Inaccuracies in the situation assessment process (level 1 and 2 SA) lead to incorrect projections
- Long- term memory does not hold relevant schemata to provide processing short-cuts
- Lack of adequate (simulation) training or clinical experience reduces exposure to decision making under pressure and makes it less likely a clinician will have treated a condition before (the "I've seen this before" feeling)
- Over-reliance on current trends presented by monitors with display characteristics which distract from relevant information leading to inaccurate projections (and interventions)

Studies undertaken in military personnel and in drivers reveal that individual SA may vary up to 10-fold.[111,112] There is, as yet, limited empirical evidence of variation in individual SA in healthcare professionals but a hypothesis extrapolating from these environments in which humans are dealing with rapidly changing conditions to healthcare is not unreasonable.

Targeted training can improve individual SA and performance (see below) but additional SA resource may always be found in other team members, it is, therefore, vital that individual SA is shared to enhance team performance.

2.2.4 TEAM SITUATION AWARENESS

It is rare for healthcare professionals to work in isolation and good quality patient care requires good teamwork.[113,114] Salas has provided a definition of the team which is useful in the context of healthcare: a team is "a distinguishable set of two or more people who interact dynamically, interdependently and adaptively toward a common and valued goal/objective/mission, who have each been assigned specific roles or functions to perform".[115] It was also highlighted in an AHRQ report in 2005 that an important characteristic of a team in healthcare is that they often function under conditions of high workload[116] placing additional constraints on the ability to achieve optimal SA.

SA is a cognitive skill and cannot, therefore, be developed by an abstract concept such as a team i.e. for a team to possess good SA each individual in the team must share their SA to ensure effective and efficient achievement of the team goal.[117] It is, therefore, likely that team SA is harder to construct and renew than individual SA.[118] However it is more unusual for a team to lose SA than a single individual [119] and the sharing of mental models between team members improves team processes such as coordination and back up and overall team performances.[120]

SA requirements for each team member will be dependent on their individual roles and sub-goals and may be conceptualised as shown in Figure 2-3.The overlapping areas in Figure 2-3 represent a requirement for shared SA so that when a task is broken down into its constituent sub-goals SA requirements may also be defined for each of those goals with clear implications for training needs in a particular context. If we consider this in the clinical setting of a cardiac arrest as described above the doctor will need to share the diagnosis of cardiac arrest and the specific rhythm the patient is in so that other team members can support the necessary steps in

managing the situation such as provision of effective chest compressions and collection of a defibrillator, which may not be in the immediate vicinity.

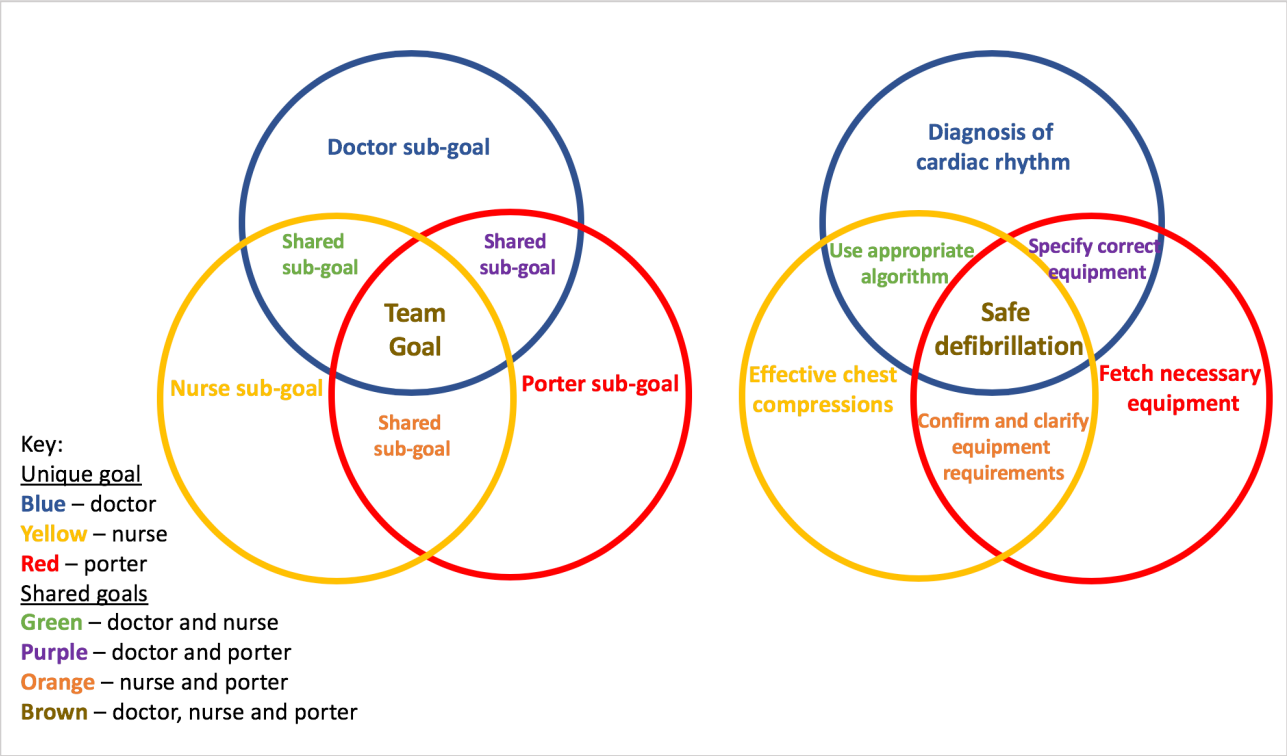


Figure 2-3: Diagram to show individual (unique) and shared sub-goals between multidisciplinary team members in a cardiac arrest situation (adapted from Endsley[121])

Research on SA has focused heavily on understanding how individual SA is developed[115] and mainly in the context of non-medical settings. Chapters 7,8 and 9 in this thesis will consider aspects of SA in teams managing simulated critical care emergencies.

2.3 CONTRIBUTION OF SITUATION AWARENESS TO ERROR

Most of the evidence considering the impact of human error in healthcare has considered NTS without distinguishing them by category. More recently evidence from a variety of work environments, including healthcare (see Chapter 1), shows that over 70% of errors involving NTS are underpinned by failings in SA[75–79,108,122,123] and that the majority of these occur at level

1.[78,108,110,122,124] Furthermore, SA errors are more common in conditions of high workload and where staff experience stress or fatigue.[125]

Vigilance, memory and judgement (the cognitive skills which underpin SA) were found to contribute to 95% of incidents involving trainee doctors and 92% in more senior clinicians in a review of 1,452 malpractice claims amongst surgical, obstetric and medical incidents.[35] Furthermore, in a study of critical incidents in anaesthesia,[78] the majority of SA errors resulted from failure to gather or perceive important information e.g. from failure to see data on a monitor obscured by surgical drapes, or misperception of information when an infusion is set up incorrectly because the two infusates look very similar. Similar results were also found in a study of death and brain injury in closed anaesthesia malpractice claims[79] and in a closed claims analysis of injury associated with monitored anaesthesia care.[126]

Errors at the level of perception are more likely where monitor interfaces are poorly designed or lack standardisation.[127] Endsley has highlighted the importance of design to support acquisition of data and enhance SA (e.g. improvements in monitor interfaces and cockpit design). However, much of the data acquired in healthcare comes from other human beings (i.e. the patient, carers and colleagues) and cannot, therefore, be improved through design but may be enhanced through training. Targeted training interventions may enhance SA and improve performance and have been designed for both individual and team SA. They can provide the techniques to enhance SA by raising awareness of what SA is and how it impacts performance, teaching strategies to improve visual search protocols, communication and coordination.[40,128–130] The accurate measurement of SA is vital in understanding the impact of training interventions and the following section will consider the available tools for the assessment of SA.

2.4 MEASUREMENT OF SITUATION AWARENESS

None of the concepts described so far is useful in describing, understanding or manipulating SA unless we can measure SA accurately and reliably. The measurement of any NTS is challenging but cognitive processes (such as SA) present particular problems because their evolution is both invisible and inaudible to the observer and, therefore, relies upon inferences drawn from actions or statements (e.g. a doctor states that the patient requires defibrillation – this would infer that she has looked at the cardiac monitor and seen an irregular rhythm [level 1 SA], diagnosed ventricular fibrillation [level 2 SA] and come to a decision that defibrillation is the appropriate action [level 3 SA]).

It is difficult to provide a truly objective measure of SA in the context of healthcare task performance. Even a superficially objective assessment of time taken for a candidate in a simulation scenario to notice a key change in a patient's condition will be subject to observer bias in determining the action or communication which indicates that the candidate is now aware of that change. Most of the available assessment tools measure groups of NTS and many of them include SA.[131] The tools for measurement of SA alone have been designed for aviation crew[132–135] and recently used in healthcare.[136,137]

2.4.1 VALIDITY AND RELIABILITY OF MEASUREMENT TOOLS

Before a measurement technique can be used for empirical research its validity and reliability must be determined. Validity refers to the extent to which the measurement is actually capturing what it sets out to measure and the overarching term, "construct validity" is now used to describe subcategories of validity.[138] The more modern approach to validity is to construct a "validity argument" that provides evidence from a range of validity domains (e.g. content and

discriminant) to support the validity of, for example, an assessment tool for the measurement of NTS in a real clinical context.[139,140]

Reliability is the accuracy with which a measurement tool will generate data under the same conditions repeatedly [141,142] and is now considered an aspect of validity.[138,142] For the purposes of this thesis, evidence of validity will be assessed using the five sources recommended by Messick[138] and described below. However, because there is no current consensus and much of the literature in medical education still uses more specific terms for validity and reliability, these are included under the relevant category:

- **Content validity:**
    - describes the extent to which a measurement tool contains all the elements that should be included to measure what it has been designed to measure it is commonly assessed by looking at the qualifications of the subject matter experts who designed the tool and the description of the steps taken to ensure that the items represent the construct being measured. Face validity is a very simple non-empirical assessment of validity e.g. “does this tool look like it would measure communication between healthcare professionals?”
- **Relations to other variables:**

  Describes the correlation of scores with other related constructs of interest e.g. do candidates with greater knowledge score more highly, or does one measure of skill in siting a nerve block compare with another? Examples of evidence describing these relationships are as follows:
    - **Discriminant validity** describes the ability of the measurement to distinguish between constructs or candidates e.g. junior trainees and senior trainees

- **Convergent validity** describes how closely aligned the measurements of one tool are in comparison to a tool designed to measure the same or different but related constructs (e.g. one would expect a tool designed to measure technical skills to give similar results to a tool for NTS in an assessment of an anaesthetist – a good candidate would be expected to have high scores in both). Evidence of convergent validity exists when two measures which should, theoretically, be related are shown to correlate strongly.
- **Concurrent validity** describes how closely aligned the measurements of one tool are in comparison to a tool designed to measure the same or similar constructs when the tools are used at the same time (e.g. two similar but distinct tests of NTS)
- **Predictive validity** describes how accurately a test predicts performance at some point in the future

- **Response process:**
  - Evidence supporting training of raters in the use of the measurement tool
  - Description and justification for the scoring system used
  - Data security and methods of reporting results
- **Internal structure:**
  - Reliability testing or factor analysis are commonly used to provide evidence for internal structure of a measurement tool. The following are examples of reliability tests
    - **Intrarater reliability** refers to the accuracy with which an individual scores the same construct using the same assessment tool on repeated occasions

        - **Interrater reliability** refers to the accuracy with which two different assessors score the same construct using the same assessment tool
        - **Internal consistency** is a measure of the correlation between different items in the same test (i.e. do items in the test measuring the same general construct produce similar scores)
        - **Test-retest reliability** is a measure of a test's ability to repeatedly replicate a result in the same situation and population
- **Consequences:**
    - Evaluation of intended or unintended consequences of an assessment e.g. evidence of unexpected bias in results
    - Methods used to determine score thresholds (i.e. pass/fail)

2.4.2 CLASSIFICATION OF MEASUREMENT TOOLS

A trainer who wishes to use an assessment tool for the measurement of NTS (including SA) must understand the benefits and pitfalls of the available techniques. Both indirect and direct, objective and subjective measurement techniques have been devised to quantify SA in individual and team contexts but it is best to consider both level of directness and objectivity as a continuum in this context, because SA is an abstract construct and, as such, cannot be measured directly. Furthermore, the objective tools which have been designed can never be absolutely free of observer bias. The majority of the objective observation tools measure SA in combination with other NTS and they will be considered generically in this chapter but in more detail in Chapter 5. Before a tool to measure SA can be designed, however, it is vital that a definition of what constitutes SA in a given situation has been determined. This assessment of SA requirements is usually undertaken by the use of questionnaires or interviews with context

specific experts and goal directed task analysis[143] to consider goals and sub goals for individuals and teams (see Figure 2-3)

Measurement techniques for SA may be categorised broadly into two domains: indirect and direct.[40,144] Table 2-2 provides an overview of tools that are relevant to or have been used in healthcare.

2.4.2.1 INDIRECT MEASURES OF SA

Indirect measurements infer SA through, for example, association of operator interactions with a system, or analysis of communication transcripts. Verbal protocols, communication analysis and psychophysiological measures (such as eye tracking) have all been used as indicators of SA in military and aviation settings. These indirect measures provide necessarily incomplete assessments of SA, for example, one cannot assume that because a participant's eyes have been directed towards an object that the object has been registered consciously (see perception and attention above) or that because a candidate moves closer to the ventilator that they have detected a change in lung compliance.

The first NTS assessment tool described for use in clinical settings was developed by Gaba and colleagues for use in the analysis of NTS in anaesthetists on a Crisis Resource Management training programme.[145] This Behaviourally Anchored Rating scale (BAR) was based on the Line Operations Safety Audit system first used in aviation and developed by Helmreich et al.[146] Since that time many NTS tools for healthcare have been designed and published in peer-reviewed journals. These tools are the subject of a systematic review which is presented in Chapter 5.

The focus of this thesis is on the measurement of SA in individuals and teams, mainly during clinical simulation training, therefore, it will focus on the assessment techniques which have been

designed, validated (sometimes in non-healthcare environments) and used in healthcare settings. Table 2-2 describes these tools and their advantages and disadvantages.

2.4.2.2 DIRECT MEASURES OF SA

The more direct assessments of SA involve measuring SA by questioning the individual and objective and subjective techniques have been devised.

2.4.2.2.1 OBJECTIVE MEASURES OF SA

Objective measures aim to assess the operator's reported SA with measures of actuality. This is more easily done in a simulated setting and requires either the use of "freeze framing" whereby the simulation scenario ceases and questions are asked regarding specific parameters after which the action resumes or "real-time questioning" where a facilitator asks questions in the moment during the simulation scenario. The Situation Awareness Global Assessment Technique (SAGAT[147]) is the most well validated of these tools and has been used in healthcare in simulated and real clinical environments.[136,148–150]

2.4.2.2.2 SUBJECTIVE MEASURES OF SITUATION AWARENESS

The development of subjective measures of situation awareness has been spearheaded by aviation educators. There is a wide variety of these tools but the Situation Awareness Rating Technique (SART)[132] is a generic tool suitable for use in other environments without adaptation. There is significant evidence of lack of reliability in self-assessment techniques with individual's commonly rating themselves more favourably than observers.[151] The Situation Awareness Rating Technique has not been used in healthcare and is described in more detail in Chapters 8 and 9.

Assessment technique	Advantages	Disadvantages
INDIRECT		
Observer-based NTS assessment tools (e.g. ANTS[152], NOTSS[153], TEAM[154])	Most commonly used tool in healthcare Some have been subject to extensive validation and reliability testing Non-intrusive Good for formative assessment	Requires extensive training Requires regular use to enhance reliability Confusing array of tools to measure NTS in a variety of settings; many poorly validated Reliability not robust enough for summative assessment (SA often cited as the most challenging domain to assess)
Physiological measures e.g. eye-tracking devices	Provides objective evidence of visual searches Relatively un-intrusive to wear (particularly when wireless)	Expensive equipment, requires extensive skill to use Can be temperamental in the field Data analysis lengthy and complicated "Look but don't see" phenomenon may reduce accuracy of data
DIRECT		
Self-assessment ratings of SA (subjective) (SART[132], SARS[135], SA_SWORD[134])	Quick and easy to administer Extensively validated Low cost Directly transferrable to healthcare setting (SART)	Problems of reliability of candidates' memory Lacks sensitivity Not yet used in healthcare
Freeze-frame, in-scenario questioning technique (objective) (SAGAT[147])	Most direct method Extensively validated Requires no conversion to healthcare setting Widely used in different settings including healthcare Data are collected in the moment	Requires simulator time Requires extensive task analysis before use Difficult to use in real clinical settings Potentially disruptive and influential on unfolding scenario

Table 2-2: Categories of tools for the indirect and direct measurement of SA with their advantages and disadvantages.

2.5 INTERVENTIONS TO IMPROVE SITUATION AWARENESS

Interventions to improve SA have been researched and developed by systems engineers, ergonomists , psychologists and educators and include: improved human-machine interfaces; adaptations to workload; properly designed alarms; appropriate automation of systems and SA-orientated training for individuals and teams.[155] An extensive review of the research into ergonomics and systems engineering approaches to improve SA is outside the remit of this thesis but it is likely that a combination of systems and training interventions will be most

effective in improving SA and more likely to lead to improvements in performance and patient outcome.

2.5.1 TRAINING INTERVENTIONS

Successful training programmes must provide participants with context specific knowledge or skills which may be transferred easily and sustainably to the workplace. There is good evidence in healthcare that team training incorporating simulation enhances performance leads to transfer of skills to the workplace[156] and improves patient outcomes.[31] David Gaba, one of the foremost pioneers of simulation training in anaesthesia, reported disappointingly high levels of failure in the management of common emergency situations (in a simulated environment) amongst residents in anaesthesia and commented: "…no industry in which human lives depend on the skilled performance of responsible operators has waited for unequivocal proof of the benefits of simulation before embracing it."[157] This worthy sentiment was appreciated by frustrated educators in healthcare. Fortunately, it is now underpinned by evidence of enhanced performance and, more importantly, improved patient outcomes after simulation training.

Simulation as applied to healthcare has been defined as "a technique, not a technology, to replace or amplify real experiences with guided experiences that evoke or replicate substantial aspects of the real world in a fully interactive manner."[158] Experiential learning is one of the key educational theories used to explain how simulation can support or enhance the transition from novice to expert professional practice. First described in ancient times by both Confucius and Aristotle and more recently by Kolb, the experiential learning cycle explores the educational psychology underpinning learning from concrete experience.[29] Even without an understanding of formal educational principles, however, it is surely self-evident that training doctors, nurses and allied health professionals together in a safe and supportive learning

environment where they can practice without the risk of doing any harm, is a good idea. This view is clearly supported by data collected from patients themselves.[159]

A recent review of team training in healthcare highlighted the upsurge in publications on the subject over the past decade. It also found that programmes (incorporating simulated practice) delivered significant benefits in team performance, medication and transfusion error and patient outcomes.[31] Interventions incorporating simulation were even more likely to produce a positive impact where they were supported by ongoing workplace initiatives after the training had taken place such as: support for staff to deliver quality improvement in the workplace; mentorship to sustain and develop leadership skills; on-going monitoring and evaluation of outcomes from training and implementation of tools designed to reinforce learning outcomes in the workplace.

Valid criticisms have been levelled at the design, implementation and analysis of team training interventions in healthcare in the past[114,160,161] but medical centres associated with the Veteran's Administration (who implemented the Patient Safety Medical Team Training programme) and hospitals that have employed the TeamSTEPPS™ system (Team Strategies and Tools to Enhance Performance and Patient Safety, sponsored by the Agency for Healthcare Research and Quality and the Department of Defence in the United States) had the ability to implement a standardised package of training across their organisations and show significant benefits in terms of patient outcomes including surgical morbidity and mortality.[162–166]

Relatively low cost training interventions to improve SA, such as video based simulation, have been shown to lead to improvements in flying and driving performance[111,167] and virtual and augmented reality training platforms are being developed in healthcare but their value as an educational tool is not yet clear.

A review of successful training techniques to improve SA for army platoon leaders, general aviation flight-crew, military pilots, and drivers has provided a list of core principles for design of these programmes which would be readily transferrable to healthcare[168]:

- Raise awareness of the construct of SA and how errors occur at each of the levels
- Teach and reinforce basic skills in finding and communicating information in the domain
- Reinforce the basic psychomotor skills necessary in the domain
- Provide training in task management and prioritisation including how to handle task interruptions, distractions and high workload situations
- Train personnel to develop higher levels of SA through explicit instruction in search strategies and communication protocols
- Practise attention sharing between key task areas and displays to improve the ability to multi-task
- Provide a broad base of experience through training in a wide variety of different situations to build schemata
- Develop team SA skills by training in cross-checking SA and sharing information across team boundaries and shift changes

2.6 CONCLUSION (WHAT THIS THESIS WILL OFFER AND FUTURE RESEARCH)

This chapter has provided an overview of the concept, origins and relevance of situation awareness in a variety of setting including healthcare. The importance of errors in SA in the evolution of critical incidents has been highlighted. A review of the tools available for measuring SA in healthcare along with techniques for training to improve SA in individuals and teams has concluded this chapter. The process of writing this chapter has revealed that there is limited

empirical evidence in healthcare on the incidence of SA in critical incidents; the tools used to measure it and techniques which may be used to improve SA in healthcare teams.

The aims of this thesis are, therefore, to provide:

- Greater understanding of error in SA in the NHS through analysis of a series of critical incidents in a large teaching hospital in the NHS (Chapters 3 and 4)
- Improved awareness of the range, validity and usability of observer-based tools for the assessment of NTS (including SA) in healthcare (Chapter 5)
- Greater clarity on the reliability and usability of observer-based tools for the assessment of NTS (including SA) (Chapter 6)
- Enhanced knowledge of the techniques of measuring SA in simulated urgent care situations (Chapters 7,8 and 9)

Finally, a review of the findings from the thesis will be presented along with the implications for training and suggestions for future research to improve SA in healthcare professionals and, ultimately, reduce error and harm in clinical settings.

CHAPTER 3 DEVELOPMENT OF A METHODOLOGY FOR ANALYSIS OF THE IMPACT OF REDUCED SITUATION AWARENESS IN SERIOUS INCIDENTS IN AN ACUTE HOSPITAL SETTING

3.1 BACKGROUND

The previous chapter has laid out the background to the requirement for a better understanding of the impact of loss of situation awareness (SA) in acute hospitals. There is limited evidence available to date on the occurrence and consequences of loss of SA in healthcare although SA errors have been cited as underpinning error in a range of healthcare settings including laparoscopic surgery[169], primary care[170] and medication administration.[171] Data from a study specifically focused on loss of SA in a database of incident reports (which were submitted voluntarily) from anaesthetists in Germany highlighted that SA was a problem in 82% of cases.[78] A further study from the same author examined the role of loss of SA in errors occurring in anaesthesia and intensive care which resulted in death and brain damage. These data were extracted from the Anesthesia Closed Claims Project database in the United States. The results showed that in 74% of cases loss of SA contributed to poor outcome.[79]

The two latter studies used Endsley's model of SA[172] to develop and operationalise a system for defining SA related errors in incidents in anaesthesia and this study has used Endsley's model to analyse errors in SA from healthcare settings across the Oxford University Hospitals NHS Foundation Trust (OUHT). The incidence and importance of loss of SA in errors occurring in an acute care trust within the NHS has not been investigated before.

Serious incidents in healthcare are almost always a combination of system failures and behavioural problems.[34,173] The importance of using an holistic approach in investigating incidents in healthcare has been highlighted by several authors.[174–176] The existing studies on the frequency of SA errors in differing contexts have shown them to be present in the chain of events leading to an error but have not considered other contributory factors and how they may be related.

The questions asked in this chapter, therefore, were designed to consider how a method of analysis might be used which incorporated the wider context of error causation in healthcare but also included a focus on the role of SA:

- There are no published methodologies for the holistic review of incident causation in healthcare – could one be devised using retrospective review of serious incidents?
- What aspects of the context in which the errors occurred (i.e. the where, when and what?) were important in the evolution of the incident?
- Could the error type (as defined by Reason's GEMS model) be defined using retrospective analysis of incidents?
- As an extension to the question above: could understanding error types in serious incidents provide further insight into the contribution of SA error?
- How can the incidence of SA error be determined in serious events in an acute hospital setting in the NHS and can it be defined by level of SA?

The next section will describe the development of the methodology for an holistic approach to understanding SA errors in acute care settings and Chapter 4 will describe the application of this method to a cohort of serious incidents from the OUHT.

3.2 OVERVIEW OF HEALTHCARE PROVISION IN THE OUHT

The OUHT is one of the largest tertiary referral teaching hospitals in the NHS and recorded over 1.4 million patient contacts, assessed 135,964 emergency department patients and delivered 7,500 babies in the year 2017-18. OUHT employs 11,612 staff, over 6,000 of whom work on the "frontline" caring for patients in both inpatient and outpatient settings, and has 48 operating theatres and seven intensive care units (data from OUHT website: https://www.ouh.nhs.uk/about/publications/documents/ouh-nhs-ft-full-accounts-2017-18.pdf). The organisation has four sites: the John Radcliffe Hospital site in Headington (incorporating the John Radcliffe itself [where all emergency admissions enter the trust and Cardiology and Cardiothoracic surgery are sited]; the West Wing [where specialist surgical services are provided]; the Children's Hospital in Oxford; the Women's Hospital [incorporating the obstetric unit and non-cancer gynaecological services] and the Oxford Eye Hospital); the Churchill Hospital on a separate site (incorporating oncology; cancer surgery; urology; dermatology and palliative care); the Nuffield Orthopaedic Hospital on another site (providing elective orthopaedic surgery; rheumatology and rehabilitation services) and the Horton Hospital in Banbury (providing district general services in North Oxfordshire, approximately 30 miles away, including a separate Emergency Department).

3.2.1 INCIDENT REPORTING AND INVESTIGATION IN THE OUHT

Since 2011 all incidents are reported in the OUHT via an electronic Datix system (Datix is a company which produces software for data acquisition and analysis in healthcare). The overarching process of reporting, investigating and reviewing serious incidents in the OUHT has undergone substantial changes since April 2015. These changes include: weekly screening of all

Datix reports by the risk management team; weekly oversight of all incident reports by the Deputy Medical Director and the Head of Governance; the institution of a weekly SIRI forum, chaired by the Deputy Medical Director, at which all serious incidents are presented in a supportive environment and a consensus is gained on the appropriate level of investigation; and a redesign of training for staff undertaking investigations including more human factors tools.

All serious incident reports are completed using the National Patient Safety Agency's guidance on reporting, in line with the recommendations in the Serious Incident Framework published by NHS England[177] (the template is included as Appendix 1). When compared with reporting rates from other acute care organisations nationally, the OUHT lies slightly above the mid-point of the cohort of all acute non-specialist NHS hospitals included in the National Reporting and Learning System (NRLS) report (https://report.nrls.nhs.uk/ExplorerTool/) as shown in Figure 3-1.

3.2.2 DEFINITION OF A SERIOUS INCIDENT REQUIRING INVESTIGATION (SIRI)

NHS England defines SIRIs as "events in healthcare where the potential for learning is so great, or the consequences to patients, families and carers, staff or organisations are so significant, that they warrant using additional resources to mount a comprehensive response. Serious incidents can extend beyond incidents which affect patients directly and include incidents which may indirectly impact patient safety or an organisation's ability to deliver ongoing healthcare."[177]

3.2.3 DEFINITION OF A NEVER EVENT

A Never Event is defined by NHS England as a, “serious, largely preventable patient safety incident that should not occur if existing national guidance or safety recommendations have been implemented by healthcare providers.”

Each year NHS England publishes the list of Never Events and in 2015/16 it included 14 incident types which are included as Appendix 2.

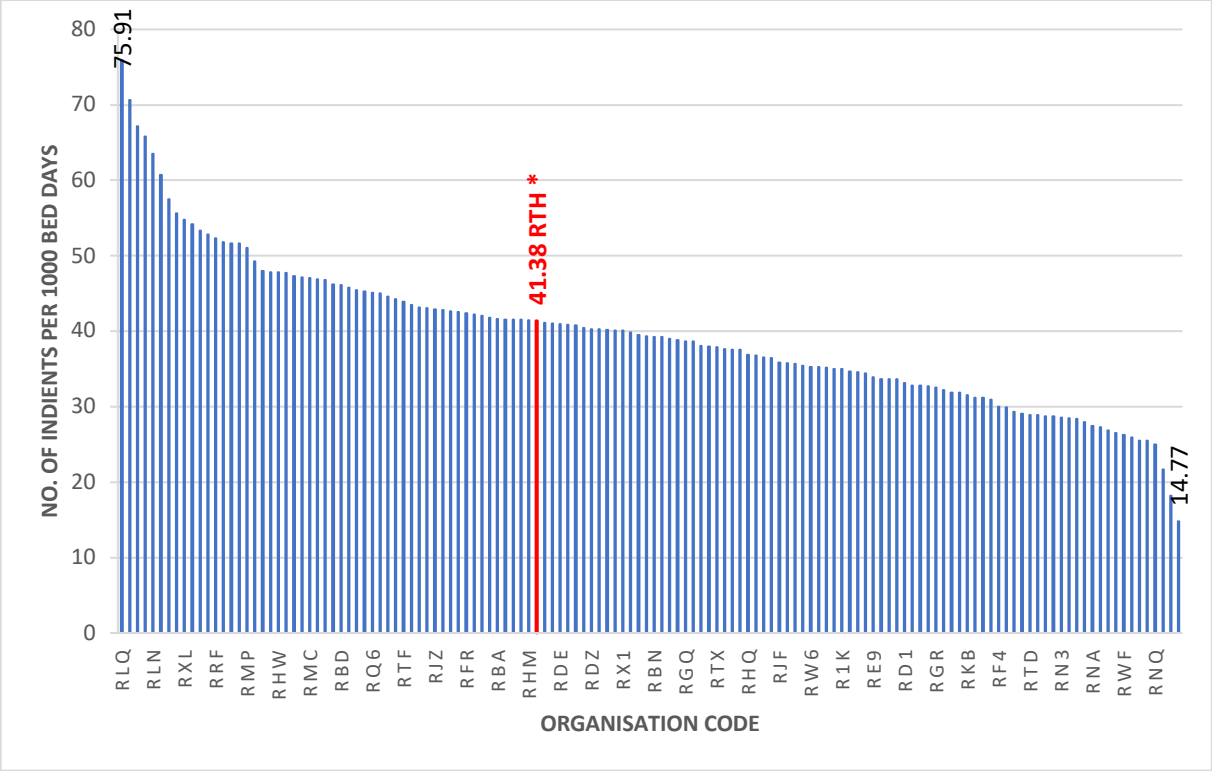


Figure 3-1: Rate of incidents reported to the NRLS per 1,000 bed days between Oct 15 and Mar 16 by all acute (non-specialist) NHS organisations . RTH*= OUHT

3.3 DEVELOPMENT OF A METHOD FOR THE ANALYSIS OF SIRIS IN THE OUHT

3.3.1 CONCEPTUAL FRAMEWORK AND STUDY PARADIGM

A pragmatic approach to the analysis of the incidents was taken and both quantitative and qualitative methods (thematic analysis) were used. The resulting methodology is reported

below and is summarised in Figure 3-4. The process was both deductive (some categories were generated prior to the analytical process) and inductive (categories emerged from the data)[178] and was informed and influenced by the professional experience of the investigators and the existing literature. Whilst the process of data analysis and interpretation was undertaken with what has been termed "empathic neutrality"[179] it is acknowledged that all research is influenced by the researcher. Repeated reflection on potential sources of bias in the context of personal beliefs and values (researcher reflexivity[179]) was a fundamental part of the process and was integral to the iterative development of the thematic analysis described below.

3.4 ANALYSIS OF ATTRIBUTES FOR ALL SIRIS: CLINICAL CONTEXT AND CONTRIBUTORY FACTORS

Following approval from the Clinical Trails and Research Governance Group Categorisation Committee at the University of Oxford this study was classified as service evaluation and registered via the OUHT Datix audit system. 17,784 incidents were reported via Datix in 2015/16 and the most serious were defined via the SIRI forum as described above.

All those put forward for investigation as SIRIs were included in the study and the reports were further analysed by the author and the Deputy Medical Director of the Trust. The incident reports were written by the healthcare professionals who had investigated the SIRIs and all investigations were supported by Divisional Governance teams who have had training in incident analysis as described above.

All incident reports were anonymised such that patients, staff and investigators were not identifiable although the professional background of those involved and the site and time of the incident were included.

Both investigators have had training and extensive experience in incident analysis in acute care settings. All data were kept on a password protected computer in a locked office or accessed via VPN from secure files held in the OUHT clinical governance shared server (this is normal policy for data protection at the OUHT and did not require modification for this study).

3.4.1 INITIAL DATA ANALYSIS

An initial analysis of all the SIRIs investigated in the OUHT in 2015/16 was undertaken by the author to gain an overview of the clinical contexts, and provide a starting point for considering the categories of information which would be collected from each incident.

Thematic analysis was used to explore and define the key attributes for the SIRIs. Thematic analysis is a qualitative research technique which has been widely used in healthcare and criteria for good thematic analysis have been described and were applied to this research[180]. The first step was a joint review and discussion of 5 SIRIs (one from each clinical division of the Trust). Data to be extracted from the complete set of incident reports were ascribed to key domains for inclusion in a table of attributes by agreement between both reviewers. A random selection of 20 SIRIs was then analysed comprehensively in isolation by both investigators with the original 5 (which had been analysed together) excluded. The purpose of this exercise was to inform further development of the list of attributes for inclusion in the analysis of all remaining incidents and to consider the role of SA error (see below). Analysis of interrater reliability for SA errors was performed at this stage in order to determine the level of agreement between both investigators. The results are presented in the next chapter.

3.4.2 DEVELOPMENT OF LIST OF INCIDENT ATTRIBUTES

The complete list of attributes for inclusion in the final analysis was achieved after the initial comprehensive review process. Definitions of attributes were informed by: the WHO International Classification for Patient Safety,[181] (which provides a comprehensive framework for classifying patient safety data and incorporates concepts developed internationally such as the JCAHO patient safety event taxonomy,[182] and the NHS National Reporting and Learning System[183]) the National Confidential Enquiry into Perioperative Death (NCEPOD) classifications (https://www.ncepod.org.uk/classification.html), the London Protocol[184] (a framework for a human factors approach to incident investigation developed in the NHS, see Appendix 3), and by the experience and knowledge of both reviewers in clinical practice (>40 years in acute care settings), education and human factors (NTS were categorised using the ANTS[152] system because this is most familiar to the author). The final attributes are shown below and the analysis of the remaining SIRIs is described in the next chapter.

During the initial process of developing the inventory of incident attributes it became apparent that some of the incidents evolved over a much shorter time frame (see below) and it seemed likely that a separate analysis of these "acute" SIRIs may be helpful, particularly in the context of designing training interventions. Consequently, all the SIRIs were analysed to consider the time from the *key error most proximal to the incident* (e.g. a decision made or action taken) until the incident happened or was discovered. These data were recorded in minutes and the results are shown in the next chapter.

3.5 ANALYSIS OF ERROR TYPE

Human error is a consistent (but not unique) feature of serious incidents in healthcare. An analysis of error type was used to provide further discrimination between incidents. There are many different published and validated taxonomies of error[56,62,185] but Professor James Reason's Generic Error Modelling System (GEMS)[21] has provided a framework for the study of errors which has been used in commercial industries[186] and in the NHS[187]. Figure 3-2 provides an overview of the GEMS algorithm. Errors were categorised at the three levels :

- Skill based ([SB] slips and lapses): a distracted doctor mistakenly starts a procedure on the incorrect side (slip); a doctor forgets to prescribe a drug for a patient to take home on discharge from hospital (lapse)
- Rule based (RB): inadequate use of good rule (swab counting procedure) leads to a swab being left inside a patient after completion of surgery
- Knowledge based (KB): in a high pressure moment a doctor misdiagnoses a cardiac arrest and delivers a DC shock when drug treatment was indicated

Reason's swiss cheese model is widely used to conceptualise the evolution of an error (the holes in the cheese link to allow an error through). I prefer, however, to think of the process (and teach it) as a chain of events because it is easier to imagine a chain being broken by an intervention, such as changing a spinal needle to one which cannot have an incorrect syringe attached or teaching safety critical communication strategies to healthcare professionals. For the purposes of this thesis the evolution of errors will be considered in this way.

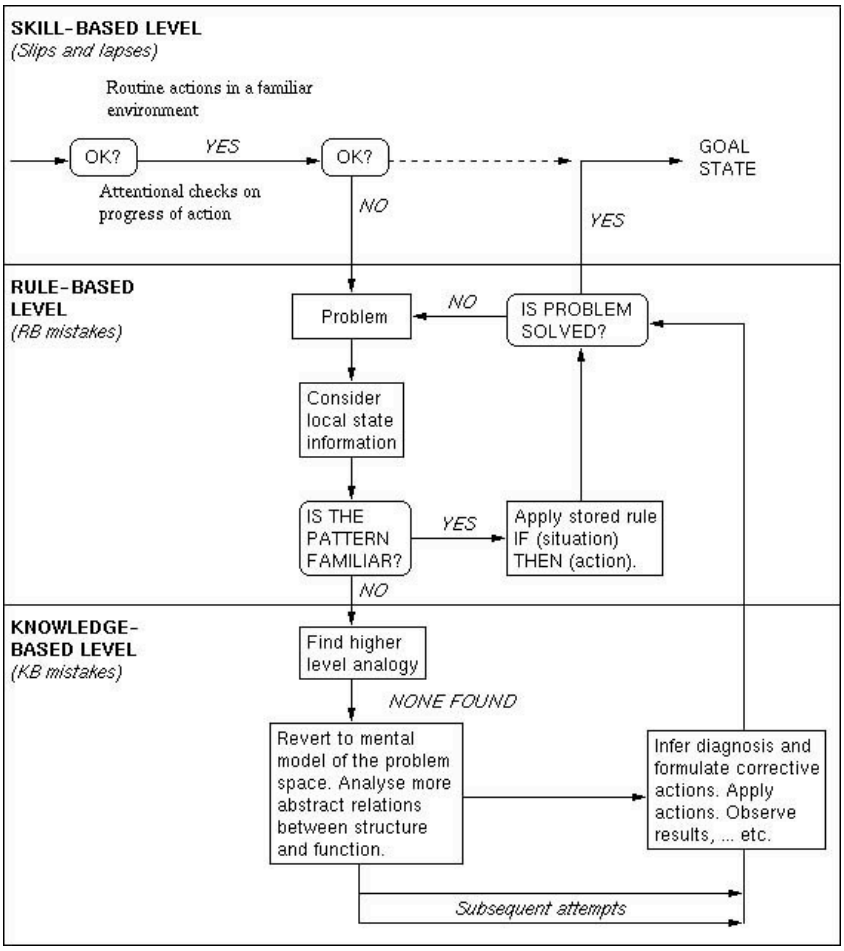


Figure 3-2: Outlining the dynamics of the Generic Error Modelling System (GEMS) reprinted from Reason J. Human Error, p64 (with permission from Cambridge University Press)

3.5.1 DEFINITION OF KEY EVENTS IN THE ERROR CHAINS

During the evolution of an incident multiple key events were usually apparent (errors in decisions or actions and system failures). Each incident was reviewed and error chains for each were defined by colour coding key events on the incident reports. The final point (i.e. the point at which GEMS error analysis would occur) was defined as that which was most proximal in time to the incident. To clarify this process an example of an error chain from the incidents in the initial analysis is shown in Figure 3-3 with key events highlighted by grey arrows and the final point with a red arrow. The error analysis was undertaken from the point of view of the primary doctor involved where possible because this is the clinical background of both investigators. In the case of incidents where other healthcare professionals were primarily

involved the error was analysed and if there was any ambiguity advice was sought from an expert in the relevant profession.

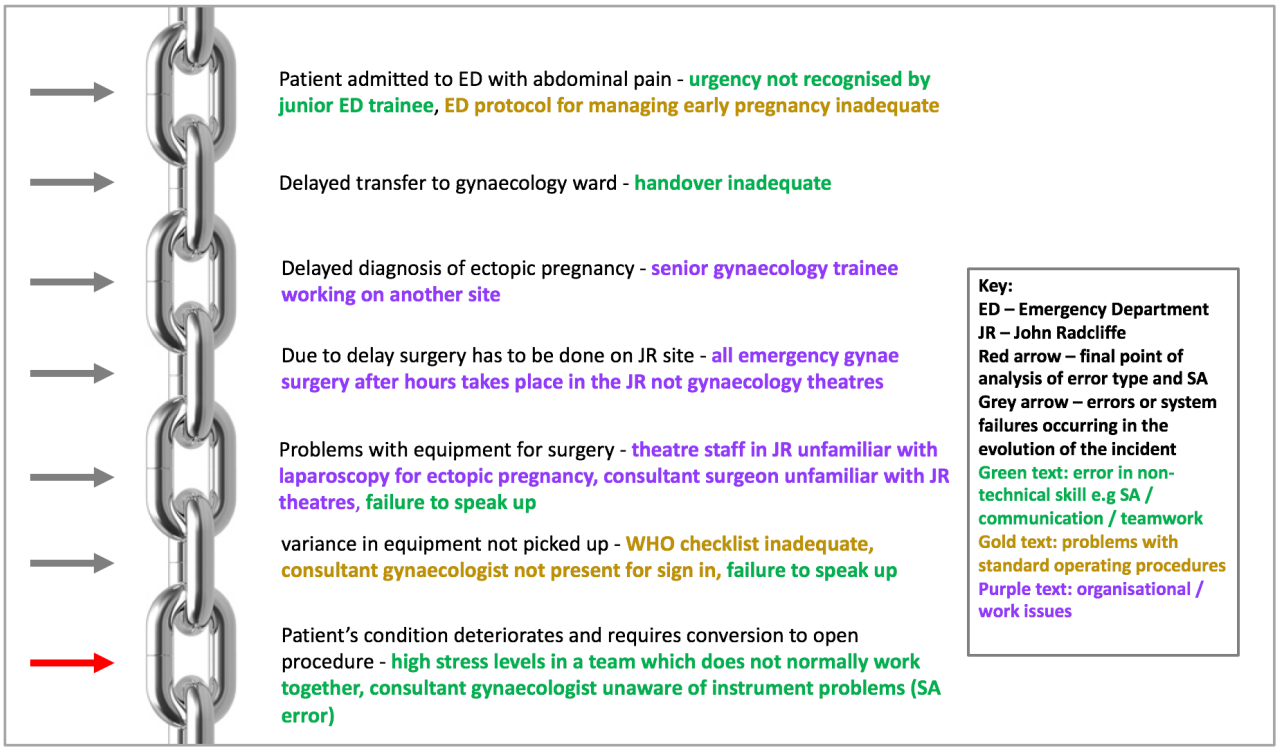


Figure 3-3: Error chain for a SIRI in gynaecology (case index no. 6). Errors occurring prior to the final key point of analysis are highlighted with grey arrows and the final error point is highlighted with a red arrow.

3.6 ANALYSIS OF SITUATION AWARENESS

Endsley's model of SA was used as the starting point for analysis in each incident and she was consulted regarding its implementation in this retrospective review of reports as she has used it in similar studies in aviation and anaesthesia.[108,188] The process of SA analysis was refined and agreed during joint analysis of the first 5 cases. Each SIRI was analysed for loss of SA at level 1,2 or 3 (including the sublevels for each category – see Table 2-1) from the viewpoint of one of the healthcare professionals involved (where possible, a doctor) at the key event most proximal to the incident occurring or to the recognition of the incident if there was a delay (see above). Where errors were detected at more than one level of SA, the lowest level was recorded. A random sample of 20 cases (as described above) were analysed by each investigator in

isolation. Inter-rater reliability was calculated at all three levels of SA with the intention of half of the remaining cases being allocated to each investigator.

3.6.1 IMPORTANCE OF SA ERROR

The initial joint review process clarified the need to cross-reference incidents with additional key attributes (shown in Table 3-1) which could help determine appropriate interventions to prevent similar errors recurring. Clearly categorising the incidents on the basis of presence or absence of SA alone would not provide the richness of detail required to understand the impact of loss of SA on the clinical situation and, more importantly, what type of training intervention (including simulation based team training) may be useful to prevent recurrence of a similar problem.

3.7 RESULTS: FINAL LIST OF INCIDENT ATTRIBUTES AND METHOD OF INCIDENT ANALYSIS

The list of incident attributes and the process for review and analysis was developed over a four month period.

The initial version of the inventory of incident attributes, after joint discussion of the first 5 cases, comprised 11 attributes this list was developed iteratively until 16 attributes were described in four categories: site and time of incident, patient factors, staff (individual and team) factors and work, environment or organisational factors. Table 3-1 shows the final list of attributes. Analysis specific to urgency of clinical condition and level of harm is further clarified below.

ATTRIBUTE	DESCRIPTION OF ATTRIBUTE
Site and Time of Incident	
Site of incident	Clinical or non-clinical area where incident occurred
Time of day	NCEPOD* classification for out of hours (18:00-07:59 weekdays and all day at weekends) and night-time (00:00-07:59 every day)
Time frame for evolution of incident	Measured in minutes from the final key error point (e.g. a decision or an action) until the incident happened or was discovered
Patient Factors	
Clinical category	Categorised according to clinical condition requiring treatment/investigation (e.g. sepsis, diabetes)
Urgency of clinical condition (see below)	NCEPOD classification(http://www.ncepod.org.uk/classification.html) for surgical cases or individual review of *medical* cases with use of context specific guidelines on urgency
Level of harm (see below)	The classification defined from each case from the SIRI forum was used – none, mild, moderate or severe
Presence of multiple comorbidities	Any additional co-morbidity (not including the presenting complaint) which was a complicating feature of the management of the case and contributed to the incident
Staff (NTS§ and Individual or Team) Factors	
Teamwork (including task allocation / management)	Problems described with the function of a team involved in the incident that include: coordination of activities; using authority and assertiveness; assessing capabilities or supporting others; appropriate task allocation; planning and preparing; prioritising tasks; providing or maintaining standards or identifying and utilising resources. The team function of exchanging information was extracted as a separate domain: "handover" and this was included in "communication"
Communication and Handover	Problems described with communication that did not relate to handover of information between teams (e.g. failure to communicate with relatives or carers, failure to communicate a change in organisational policy) or problems with handover of information between teams or individuals involved in the incident
Decision making	Problems described with identifying options, balancing risks, re-evaluating the situation
Individual (staff) Stress / fatigue	Stress and/or fatigue that were described as contributory factors in the incident
Distraction	Distraction of a member of staff described as a contributory factor in the incident
Work / Environment or Organisational Factors	
Work / Environment (staffing levels)	Inadequate staffing levels cited as a factor in the incident
Work / Environment (Equipment or Technology)	Problems described with: mechanical or electronic devices and poor equipment design (including documentation systems), availability or utilisation
Organisational (Electronic patient record [EPR])	Problems described with OUHT's EPR system: this was considered separately from equipment and technology as the OUHT has a bespoke EPR system which presents unique issues with prescribing and recording of patient information
Task / Technology (Availability and use of protocols)	Poor equipment design or problems with use of protocols (Standard Operating Procedures) e.g. used inappropriately, not used, or not available or fit for purpose.

Table 3-1: Final list of incident attributes for categorisation of SIRIs in the OUHT (divided into incident site and time, patient, staff and work domains) . *NCEPOD: National Confidential Enquiry into Perioperative Death. §NTS categorised using ANTS taxonomy

3.7.1 PATIENT FACTORS - URGENCY OF CLINICAL CONDITION

Each surgical SIRI was given an acuity code in line with NCEPOD's guidance on urgency of care which defines four categories: elective, expedited, urgent and immediate (full description in Appendix 4). The NCEPOD classification was developed to categorise patients undergoing surgical procedures but was adapted for the medical cases in this study to enable the same system to be used for both.

Patients with medical conditions not requiring surgery are also subject to similar considerations on the grounds of acuity of their illness for example, guidance on the management of acute coronary syndromes specifies time points for pharmacological and/or invasive interventions (e.g. angioplasty or coronary stenting using NICE guidance on the management of acute coronary syndrome) and subsequent allocation to NCEPOD categories. If appropriate guidelines were not available expert advice was sought and a consensus view on urgency attained. For the non-acute SIRIs in which the care of the patient required them to be in hospital for the management of a chronic condition (e.g. dementia or palliation of a cancer) the acuity code was allocated according to the level of medical input required. Examples of medical cases allocated to each of the acuity codes are given in Appendix 5 in the following chapter.

3.7.2 PATIENT FACTORS - LEVEL OF HARM

Level of harm was categorised according to the definitions provided by the NRLS (see Table 3-2) and described with examples in the following chapter. For the purposes of this study we combined severe harm and death and analysed them as one category.

Definition of level of harm
0 – no harm
1 - low level of harm: any unexpected or unintended incident that required extra observation or minor treatment and caused minimal harm to one or more persons
2 - moderate harm: any unexpected or unintended incident that resulted in further treatment, possible surgical intervention, cancelling of treatment, or transfer to another area, and which caused short-term harm to one or more persons
3 – severe harm or death: any unexpected or unintended incident that caused permanent or long-term damage or death to one or more persons

Table 3-2: Definition of level of harm according the National Reporting and Learning System

3.7.3 FINALISED METHOD OF ANALYSIS FOR 167 SIRIS

The method of analysis described above was then applied to the full cohort of SIRIs from 2015-16. The process is summarised in Figure 3-4. Results of the analysis are presented in the next chapter.

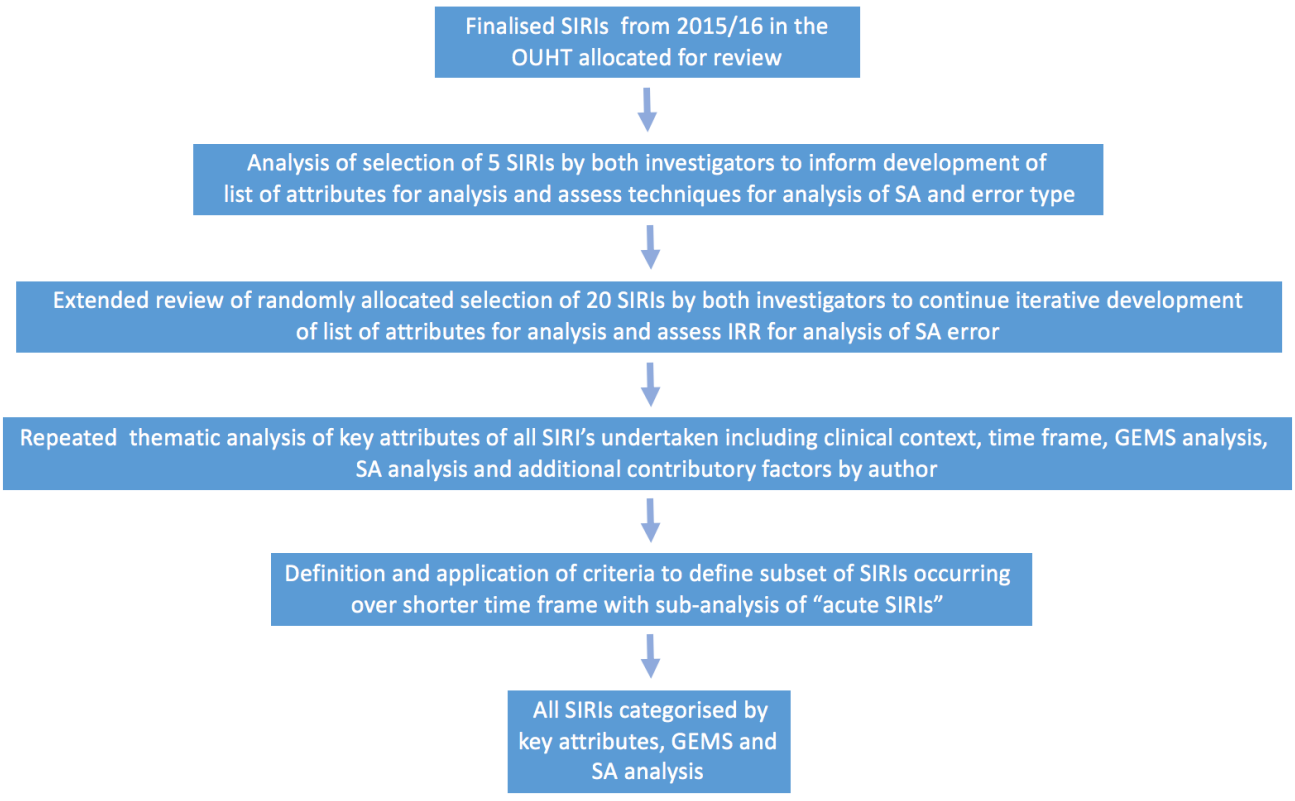


Figure 3-4: Overview of flow chart describing method of analysis of SIRIs in the OUHT and applied to 167 SIRIs from 2105-16; IRR: interrater reliability.

3.8 DISCUSSION

The method devised in this chapter has provided an holistic framework for the review of incidents in an acute care settings in the NHS. The two investigators involved (the author and the deputy medical director) took four months to devise the final list of attributes and the technique for analysis of error and SA (including the calculation of IRR). The list of incident attributes may require refinement for use in other similar organisations (e.g. hospitals where EPR is not used) but the method described would be straightforward to apply. However, understanding the necessity for relevant subject matter expertise and time resources would be vital.

3.9 CONCLUSION

This chapter has described the development of an holistic method of analysing critical incidents in an acute care hospital in the NHS. A list of key attributes incorporating systems problems and behavioural errors was produced and combined with analysis of error type and SA errors.

The next chapter will describe the application of the method and results of the analysis for all the SIRIs in the OUHT in 2015-16.

CHAPTER 4 THEMATIC ANALYSIS OF THE IMPACT OF REDUCED SITUATION AWARENESS IN 167 SERIOUS INCIDENTS REQUIRING INVESTIGATION IN AN ACUTE HOSPITAL SETTING

4.1 INTRODUCTION

The previous chapter has described the development of an holistic method for the retrospective analysis of the impact of SA error in critical incidents in an acute care hospital in the NHS (the OUHT). This chapter will describe the application of that method to the investigation of 167 SIRIs which occurred in the OUHT in 2015-16. The questions to be answered in this study are as follows:

- What were the attributes (the where, when, and what) of the SIRIs occurring in the OUHT?
- What types of error were being made by healthcare professionals and could this provide further insight into the contribution of SA error in critical incidents?
- What was the incidence of SA error in serious events in an acute hospital setting?
- What was the distribution of error within the 3 levels of SA?
- What was the impact of loss of SA in a subgroup of incidents where incidents evolved over a short time frame?

4.2 METHOD

179 SIRIs were declared to the Oxfordshire Clinical Commissioning Group (OCCG) during the financial year 2015/16, 12 of which were subsequently downgraded after further investigation and with agreement from the OCCG. Therefore, 167 SIRI reports were analysed.

Five were analysed by both investigators with a further 20 randomised and reviewed separately as described in the previous chapter.

Subsequently incident analysis was undertaken iteratively by the author alone and any ambiguous decisions were discussed and agreed together. SIRIs were categorised according to the method described in the previous chapter, by incident attributes (summarised for convenience in Table 4-1), GEMS error analysis and SA error. A flow chart describing the method of analysis is shown in Figure 4-1 .

ATTRIBUTE
Site and Time of Incident
Site of incident
Time of day
Time frame for evolution of incident
Patient Factors
Clinical category
Urgency of clinical condition
Level of harm
Presence of multiple comorbidities
Staff (Individual or Team) Factors
Teamwork including task allocation / management
Communication and handover
Decision making
Individual (staff) Stress / fatigue
Distraction
Work / Environment or Organisational Factors
Work / Environment (staffing levels)
Work / Environment (Equipment or Technology)
Organisational (Electronic patient record [EPR])
Task / Technology (Availability and use of protocols)

Table 4-1: Summary list of finalised incident attributes for 167 SIRIs in the OUHT between 2015-16

4.2.1 ANALYSIS OF A SUBSET OF ACUTE SIRIS

The acute SIRIs were separated from the whole cohort by analysis of the time frame over which the incident evolved. Cases for analysis in this subset were excluded if:

- They did not involve direct patient care
- They did not occur in an inpatient setting (e.g. happened in outpatients or were administrative errors)
- The incident evolved over a time frame of greater than 8 hours (this time was chosen arbitrarily in the first instance because it represents the recommended length of a shift for doctors in training)

4.3 DATA ANALYSIS

Frequency of occurrence was calculated for all incident attributes in both the acute and non-acute SIRIs. Where possible all attributes were coded in binary terms as present or absent (e.g. presence of any communication error was coded as “1” for present and “0” for absent regardless of whether this was a handover problem or failure to write to a patient with a request to attend a clinic). All data (both quantitative and qualitative) were stored on a university computer in a locked office.

Odds ratios were calculated to determine the likelihood of an acute SIRI happening out of hours, skills, rules or knowledge based error occurring in an acute or non-acute SIRI and the likelihood of SA error in acute or non-acute SIRIs. All statistical analyses were calculated in SPSS (IBM® V24.0).

4.4 RESULTS

The reports averaged 16 pages in length (maximum 37 pages and minimum 6 pages) and a total of 2655 pages were analysed iteratively by the author. The process of analysis is summarised in Figure 4-1.

There were seven never events: four wrong site surgeries, two wrong site blocks and one swab left in a wound. Three of the reports referred to more than one incident. The first was a group of 5 incidents in ophthalmology where patients had either received the wrong medication injected into their eye or the wrong eye had been injected. Only two of these cases presented an appropriate level of detail (these were the most recent cases) the remainder had occurred over 10 months earlier and had come to light after the latter two were investigated. In this case the two most recent cases were included in the study (case index nos. 44 and 45) but the other

3 were not. The second incident report covered two fully investigated cases of pressure ulcers developing at the same time on a medical ward (case index nos. 73 and 74) both were comprehensive reports and so were included as two separate cases. The third report referred to an unfortunate patient who suffered two falls on different wards during her protracted stay in the hospital (case index nos. 79 and 80). Both incidents were comprehensively investigated and were reported separately on Datix, therefore, they were analysed individually.

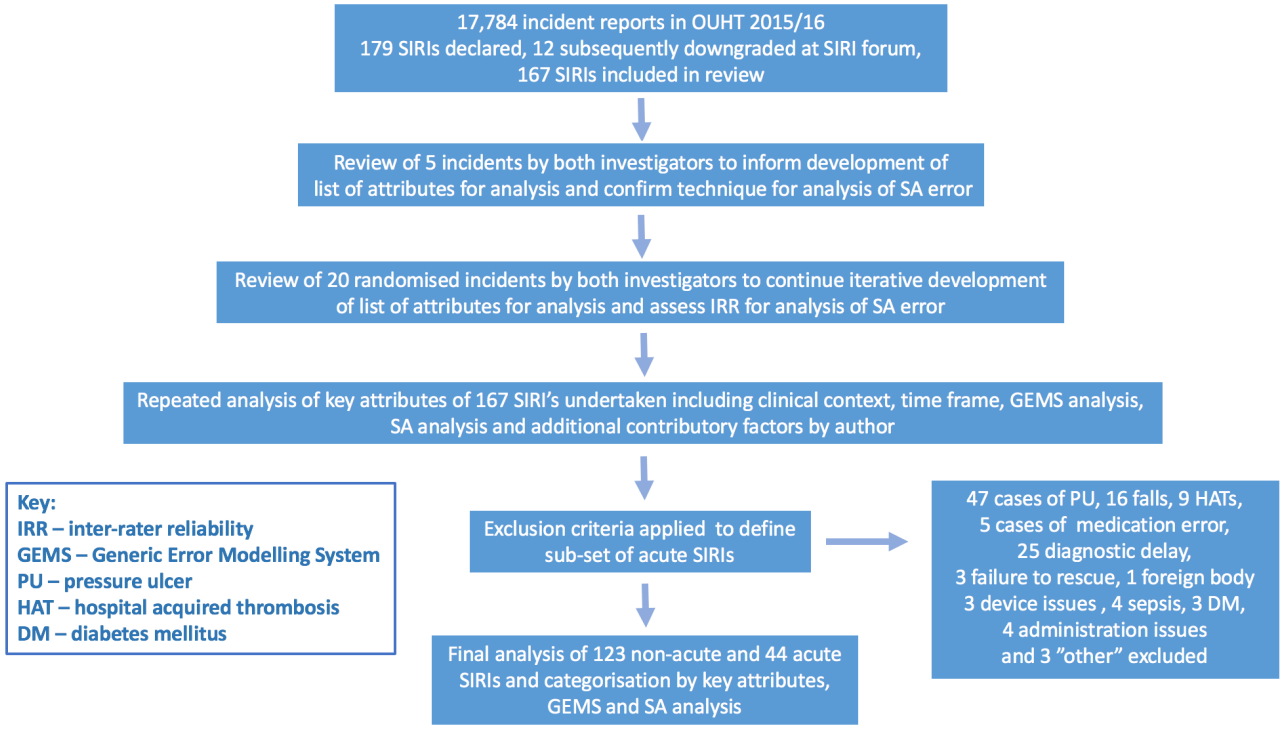


Figure 4-1: Flow chart describing detail of method of analysis of 167 SIRIs occurring in the OUHT in 2015-16

4.5 INCIDENT ATTRIBUTES

All SIRIs were categorised according to the 16 attributes outlined in Table 4-1. Table 4-2 shows the incidence of each of the attributes for non-acute and acute SIRIs in the patient, staff and work categories (these are explored in more depth later in this section). The site and timings of the incidents and the clinical categories are described separately below.

Incident Attribute	Occurrence in non-acute SIRIs (total 123) no. (%)	Occurrence in acute care SIRIs (total 44) no. (%)
Patient Factors		
Urgency of clinical condition*	34 (28)	33 (75)
Level of harm§	29 (24)	9 (20)
Presence of multiple comorbidities	58 (47)	9 (20)
Staff (NTS, individual or team) factors		
Teamwork error	20 (16)	21 (48)
Communication error	88 (72)	36 (78)
Decision making error	82 (67)	25 (57)
Stress / fatigue	6 (49)	12 (27)
Distraction	33 (27)	18 (41)
Work / Environment or Organisational Factors		
Staffing levels	41 (33)	14 (30)
Equipment or technology	48 (39)	26 (59)
EPR	35 (28)	4 (9)
Availability / use of protocols	111 (90)	38 (86)

Table 4-2: Incidence of attributes from patient, staff and work domains in non-acute and acute SIRIs. * urgent and emergency cases; § severe harm or death

4.5.1 SITE OF INCIDENT

Figure 4-2 shows the sites of all 167 SIRIs. The majority of SIRIs in the OUHT occurred in ward settings (62%). The acute SIRIs occurred more frequently in Theatres, ICU, ED and Maternity (61%) than non-acute SIRIs (1%). Incidents in outpatients included administrative problems, delays in diagnosis and one fall. The majority of incidents radiology (5 [71%]) were diagnostic delays (case index nos. 49,67,68,113,143) one case involved administration of the incorrect radiopharmaceutical agent to a child (case index no. 101) and the other was a never event (a procedure to remove a portacath (a long central-venous line) was begun on the incorrect side, case index no. 78). The two cases in ambulatory medicine were failure to follow up a patient with pneumonia who was receiving IV antibiotics from the ambulatory care ward (case index no. 146) and the other was an incorrect decision to manage a patient with type 2 diabetes who presented with high blood sugar levels as an outpatient (case index no. 153).

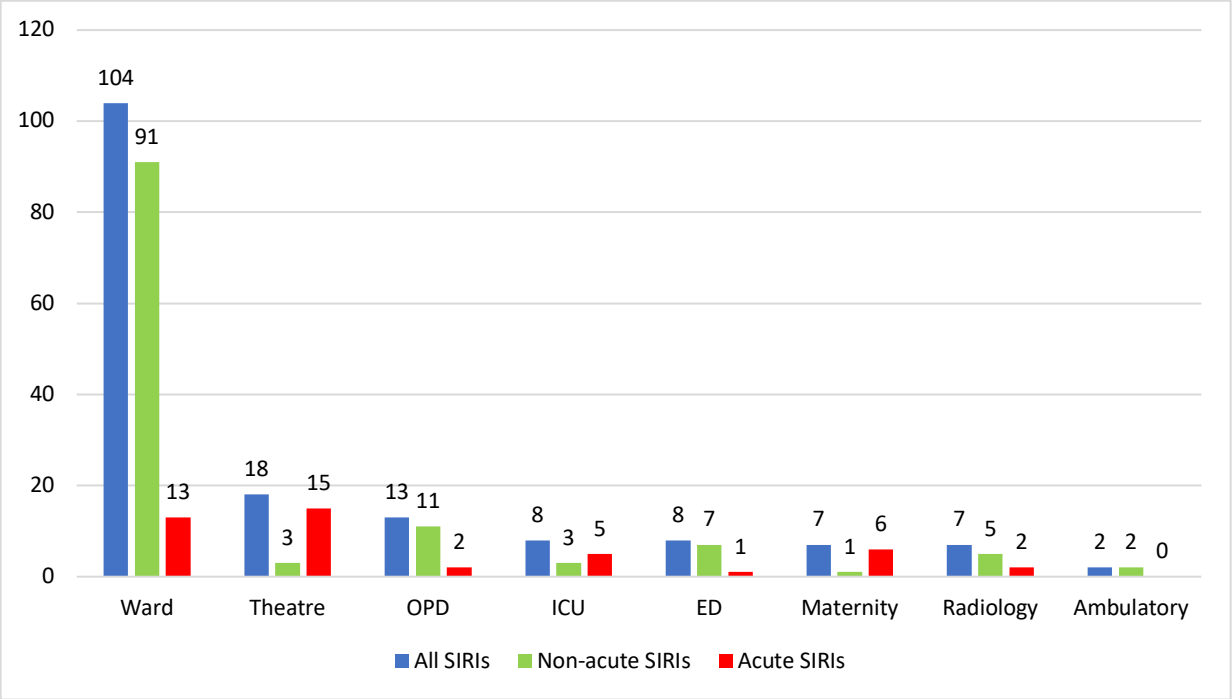


Figure 4-2: Site of occurrence of 167 SIRIs in the OUHT (OPD: Outpatients Department; ICU: Intensive Care Unit; ED: Emergency Department)

4.5.2 TIME OF DAY

The time of day was classified according to the NCEPOD categorisation of out of hours work (18:00-07:59 weekdays and all day at weekends and night-time [00:00-07:59 every day]). Overall, 32 (19%) incidents occurred out of hours. Within the acute SIRIs 19 (43%) occurred out of hours and 13 (10%) of non-acute SIRIs happened out of hours. The odds ratio (with 95% C.I.) for the likelihood of occurrence of an acute SIRI out of hours was 6.43 (2.81,14.72).

4.5.3 TIME FRAME FOR EVOLUTION OF INCIDENT

The chain of error described above was analysed for each incident and the time calculated from the point of the final error until the incident occurred or was detected. Any ambiguity on time of final error to incident was discussed and agreed by both investigators.

Most of the acute SIRIs had an estimate of time taken included in the report but they occasionally required a sensible estimate of time scale for example: a patient was undergoing an elective robotic resection of his prostate and an instrument was incorrectly attached to the robot and damaged one of the blood vessels in the pelvis. There is no recorded exact time from the point of insertion of the instrument onto the robotic arm (the final error point in this case) and damage to the blood vessel but a considered estimate of 5 minutes was made and agreed by both investigators (case index no. 18).

In some of the non-acute cases a considered estimate of time frame also had to be made for example: a patient admitted to a ward and developing a pressure ulcer is an incident which occurs over a period of days and most of the errors in these cases were lapses in checking the skin or failure to follow the standard operating procedures for assessment of risk of pressure ulceration (or both). The final error point in these cases was taken from the final recorded episode of skin review or reassessment. Furthermore, in some of the cases of delayed diagnosis the initial investigation (often a scan) had been undertaken years previously. If the patient had been reviewed in clinical settings during the time since the initial investigation, and it may reasonably have been expected that their previous scans would be reviewed, then the time was calculated from the point of that interaction with a member of the clinical team. If no such interaction had occurred then the time was calculated from the point of the initial scan to the time the SIRI was detected.

Figure 4-3 shows the time frames for acute and non-acute SIRI. The average length of time for incident evolution in the acute SIRIs was 80 minutes (mode 30 minutes, median 30 minutes) and for the non-acute SIRIs 104217 minutes (mode 7200 minutes, median 8640 minutes). The cut off for categorising incidents as acute or non-acute was made at 390 minutes (6.5 hours)

because there was then a hiatus in incident occurrence until 495 minutes (8.25 hours). There was an incident recorded which had a time frame of 472 minutes but this was a case of a fall on the surgical emergency unit in a patient who did not receive a timely assessment of falls risk but would not have been found to be at risk of falls even if the assessment had been complete. It was found to be an unavoidable accident and no error has been ascribed.

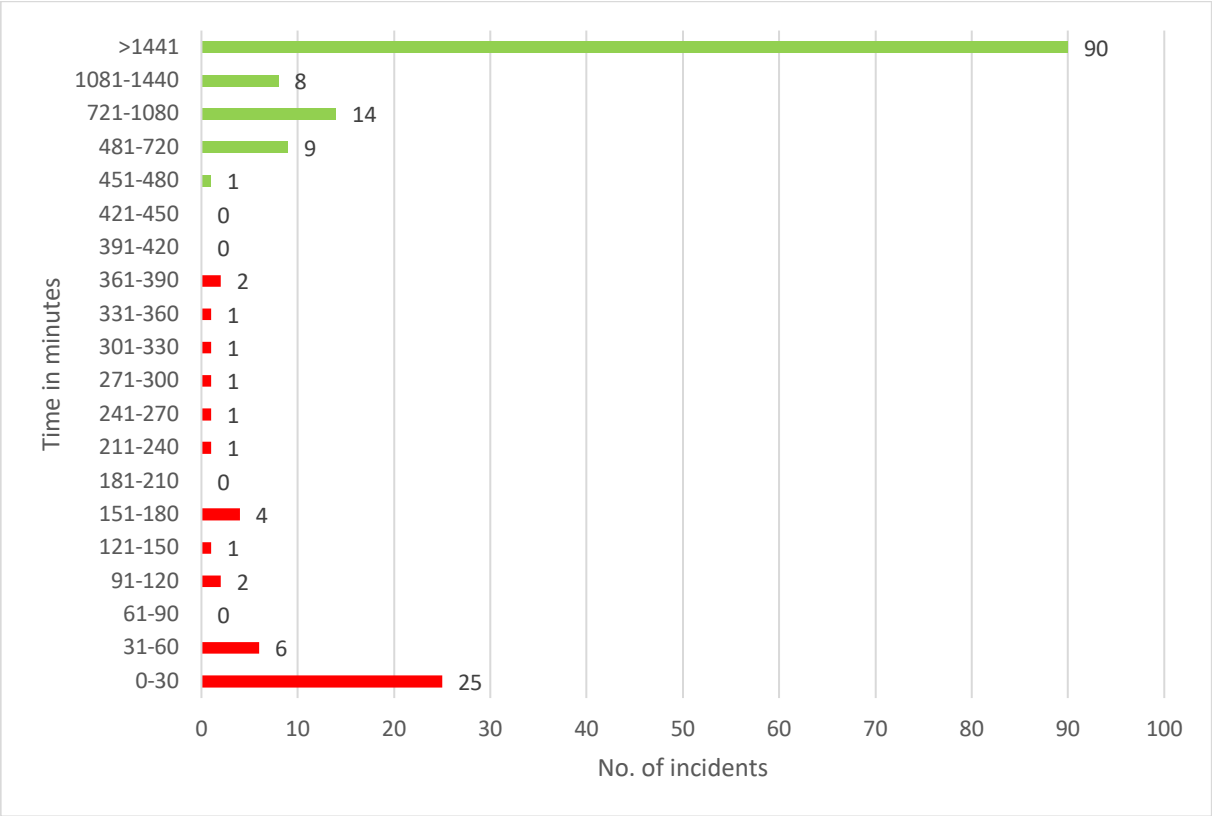


Figure 4-3: Time frame (in minutes) for evolution of 167 SIRIs. Acute SIRIs in red non-acute SIRIs in green.

4.5.4 PATIENT FACTORS

There were four patient factors in the final list of attributes and they were analysed using a mixed methods approach as described below.

4.5.4.1 CLINICAL CATEGORIES FOR 167 SIRIS

The SIRI's were categorised into thirteen clinical themes which were aligned as closely as possible with local and national NHS quality priorities for 2015/16. Figure 4-4 shows incidents classified by clinical category for all SIRIs with a separate column for non-acute and acute SIRIs highlighted in green and red respectively.

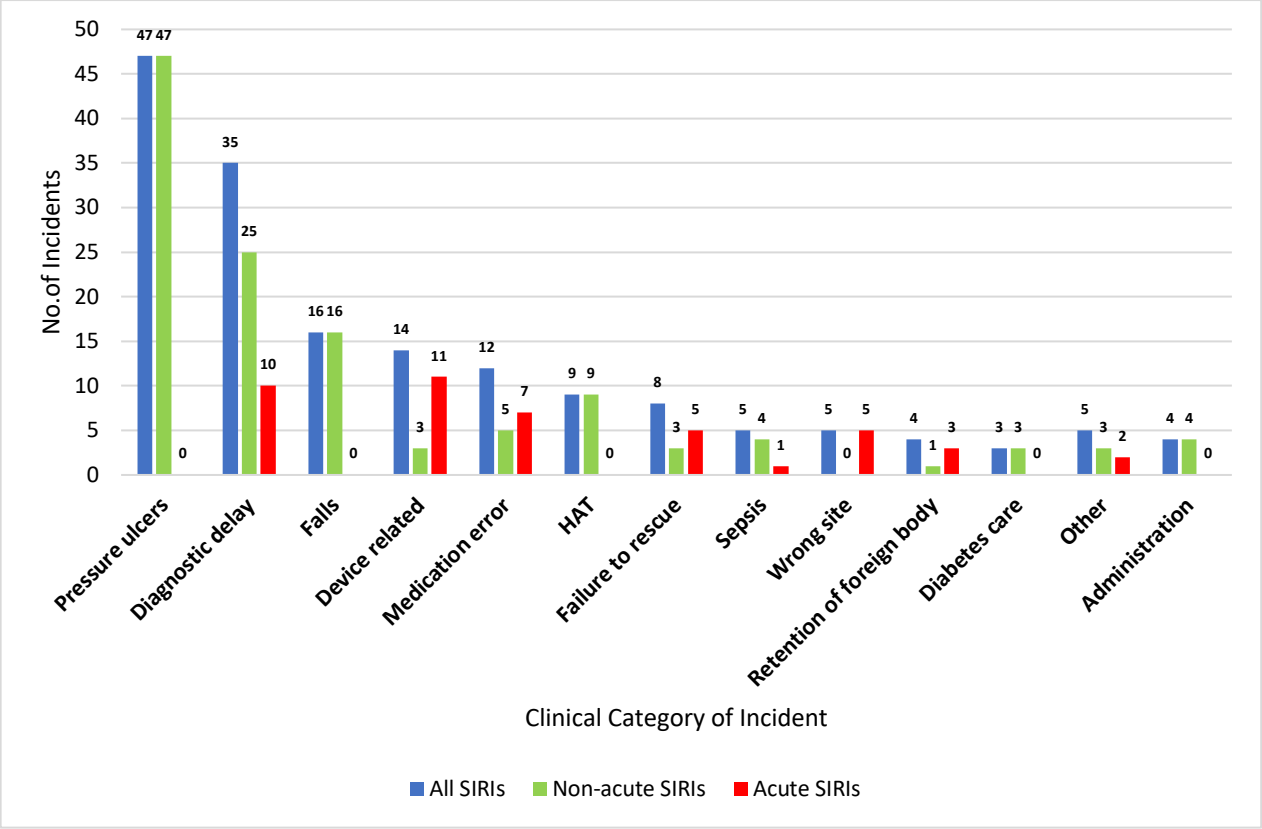


Figure 4-4: Clinical categories of 167 SIRIs with subdivision into non-acute and acute SIRIs (HAT – hospital acquired thrombosis)

Only 9 SIRIs (categorised as "administration" and "other") did not fit into the other themes. The 4 cases of administrative error included: letters not sent from ophthalmology outpatients; problems with adequate date entry into a national database for monitoring of abdominal aortic aneurysms; a letter for another patient inadvertently filed in the incorrect set of notes and a

patient with oral cancer who was inadvertently lost to follow up after wrongly being cancelled from a clinic.

The 5 cases which were allocated to the "other" category included: a fire in cardiothoracic theatres; a power failure to all theatres on the Churchill site; the accidental use of unsterile equipment for transplant surgery; a patient who absconded from the Emergency Department and later committed suicide and mis-labelling of blood samples in outpatients requiring the patients to return for re-sampling.

13 incidents were initially categorised in two or more domains, further analysis to define the overarching categorisation is described in Appendix 6.

4.5.4.2 URGENCY OF CLINICAL CONDITION

The urgency of the patient's clinical condition was assessed at the time of final error point (see example error chains in Figure 4-9 and Figure 4-10) and an NCEPOD category was allocated as described in the previous chapter.

Three incidents were excluded from this analysis because the acuity of the patients' conditions was not felt to be relevant: the first (case index no. 32) was an error which was the result of a software problem affecting 71 patients in an abdominal aortic aneurysm screening programme, the acuity of the patients' clinical condition was not a factor in the problems experienced with the human/machine interface; the second (case index no. 11) was an incident relating to a power failure at the Churchill Hospital theatres as a result of inadequate servicing protocols being in place, several patients were affected on the day but consideration of acuity of patient care at the point of error (which had occurred months before the incident) was not possible and neither was it felt to be a relevant consideration for the case because it had no bearing on

the decision making of the company involved; the third (case index 38) involved a patient visiting the outpatient department and tripping on an uneven floor outside the hospital canteen on the way home, the acuity of the patient's clinical condition was not relevant to the error on the part of the estates team in not repairing the surface.

The acuity of each clinical condition at the final error point was considered for the remaining 164 incidents (see Figure 4-5).

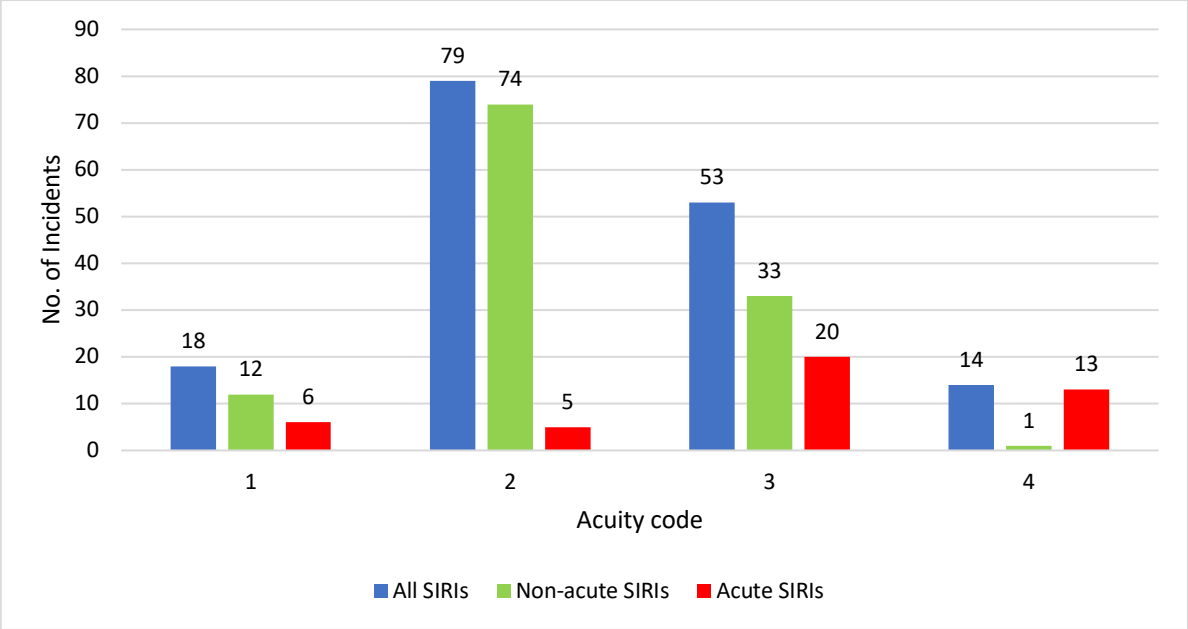


Figure 4-5: Acuity code, as defined by NCEPOD, for 164 SIRIs; 3 incidents were excluded because the acuity code was not relevant (see explanation in section 4.5.4.2): 1-elective, 2-expedited, 3-urgent, 4-immediate

Incidents involving patients categorised as NCEPOD "urgent" and "emergency" cases were more common in the acute SIRIs (75%) than the non-acute SIRIs (28%).

4.5.4.3 LEVEL OF HARM

Level of harm was categorised according to the definitions provided by the NRLS and described with examples in Table 4-3. One incident report of three patients who were lost to follow up from the eye clinic was excluded from the analysis as the patients were still being followed up (case index no. 20 in the non-acute SIRIs).

29 (24%) of the non-acute SIRIs caused severe harm, including death; 85 (69%) cases caused moderate harm; 7 (6%) caused minor harm and 1 (1%) caused no harm. Pressure ulceration is classed as a SIRI at Category 3 (full thickness skin loss) and these formed the largest group of non-acute SIRIs. This type of ulcer is automatically categorised as moderate harm.

Nine (20%) of the acute SIRIs caused severe harm, 23 (52%) caused moderate harm; 10 (23%) caused minor harm and 2 (5%) caused no harm (see Figure 4-6).

Definition of level of harm	Example from OUHT SIRIs (case index no.)
0 – no harm	A patient returned to the intensive care unit after a long operation and one of the bite blocks which had been sited at the beginning of the case was discovered in the patient's mouth – it had inadvertently been left in place but had caused no harm (50)
1 - low level of harm: any unexpected or unintended incident that required extra observation or minor treatment and caused minimal harm to one or more persons	A patient was admitted electively for the removal of a long central line (portacath) after treatment was completed. The procedure was begun on the wrong side of the chest. The problem was promptly recognised and the procedure completed on the correct side with minimal increase in the length of procedure, and no increased length of stay (78)
2 - moderate harm: any unexpected or unintended incident that resulted in further treatment, possible surgical intervention, cancelling of treatment, or transfer to another area, and which caused short-term harm to one or more persons	A patient developed a category 3 pressure ulcer on the inner aspect of the knee underneath an anti-embolism stocking (102)
3 – severe harm or death: any unexpected or unintended incident that caused permanent or long-term damage or death to one or more persons	Severe harm: A patient admitted for investigation of weight loss and anaemia had a scan which recommended additional investigations which were not organised, leading to a 2 year delay in the diagnosis of primary liver cancer which was then untreatable (132) Death: A patient had an emergency laparotomy for a perforated duodenal ulcer which was successfully repaired. He was discharged home without the necessary medication to protect his stomach from further similar problems and he returned 2 weeks later with another catastrophic bleed which unfortunately led to his death (127)

Table 4-3: Description of level of harm (according to definitions from NRLS) to patients with example incidents from 167 SIRI's in the OUHT between 2015-16

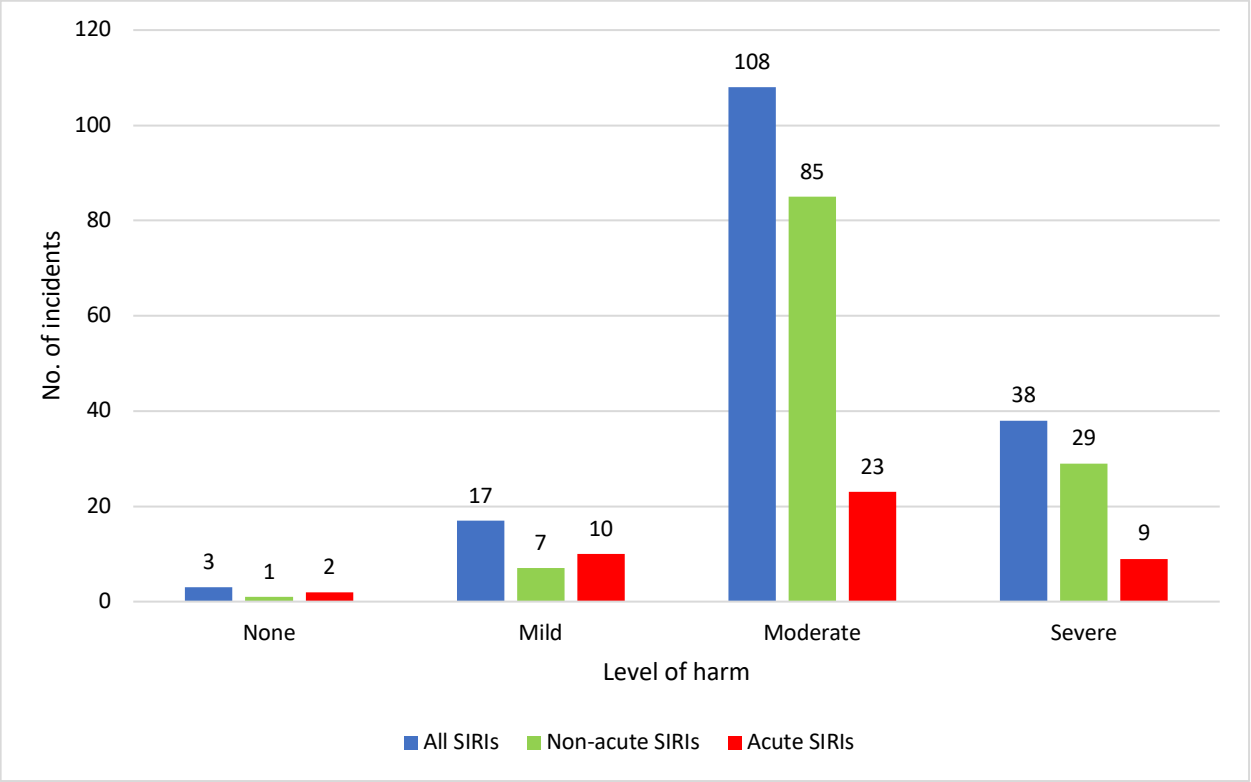


Figure 4-6: Level of harm to patients categorised as described in table 4-3 (severe includes death) for all SIRIs and subdivisions of acute and non-acute SIRIs

4.5.4.4 PRESENCE OF MULTIPLE COMORBIDITIES

Each incident was analysed to consider whether the presence of additional comorbidities provided further challenges to the patient's management and contributed to the evolution of the incident. Examples include:

- a patient with a complicated past medical history including diabetes, ischaemic heart disease, peripheral vascular disease and gout was admitted with sepsis secondary to an infected foot. He had surgery promptly to amputate some toes and in the postoperative period he suffered a heart attack which was "silent" because his diabetes had damaged the autonomic nerve supply of his heart making it more difficult for staff to diagnose the problem. (Case index no. 133)

- a patient with diabetes was admitted having fallen and suffered a subdural haematoma. He had surgery to remove the clot from his brain but postoperatively his continued confusion was ascribed to the head injury and the diabetic ketoacidosis was missed. (Case index no. 99)
- many of the cases of pressure ulceration occurred in patients with pre-existing dementia or where acute delirium complicated their care (through difficulty in communication and lack of compliance with care) and contributed to the development of skin breakdown.

The presence of comorbidities contributed to the evolution of the incident in 58 (47%) of non-acute SIRIs and in 9 (20%) of the acute SIRI's.

4.6 STAFF (NTS, INDIVIDUAL OR TEAM) FACTORS

Figure 4-7 shows the incidence of problems with teamwork, communication, decision making, stress or fatigue and distraction described in the reports.

Teamwork errors were more common in the acute SIRIs than the non-acute SIRIs (48% vs 16% respectively) and the majority of these were either task management or allocation problems (which are shown separately in Figure 4-7).

Communication problems were common (76% of all SIRIs) and were similarly common in acute and non-acute SIRIs (82% vs 71%) Initially communication and handover had been considered separately but they were combined because in the 124 cases of communication errors across all SIRIs 98 (79%) were caused by handover together with other communication issues (e.g. poor communication between surgeons and anaesthetist leading to delays and heightened tension in the anaesthetic room where a block was then sited on the incorrect side[case index

no. 144]) and in no case was handover an isolated communication problem. In 22% of cases communication was a problem in isolation (e.g. failure to send follow up letters from outpatient settings to patients).

Stress and fatigue were more commonly described in non-acute SIRIs (49%) than acute SIRIs (27%) but the reverse was true for distraction (acute SIRIs, 41% and non-acute SIRIs, 27%).

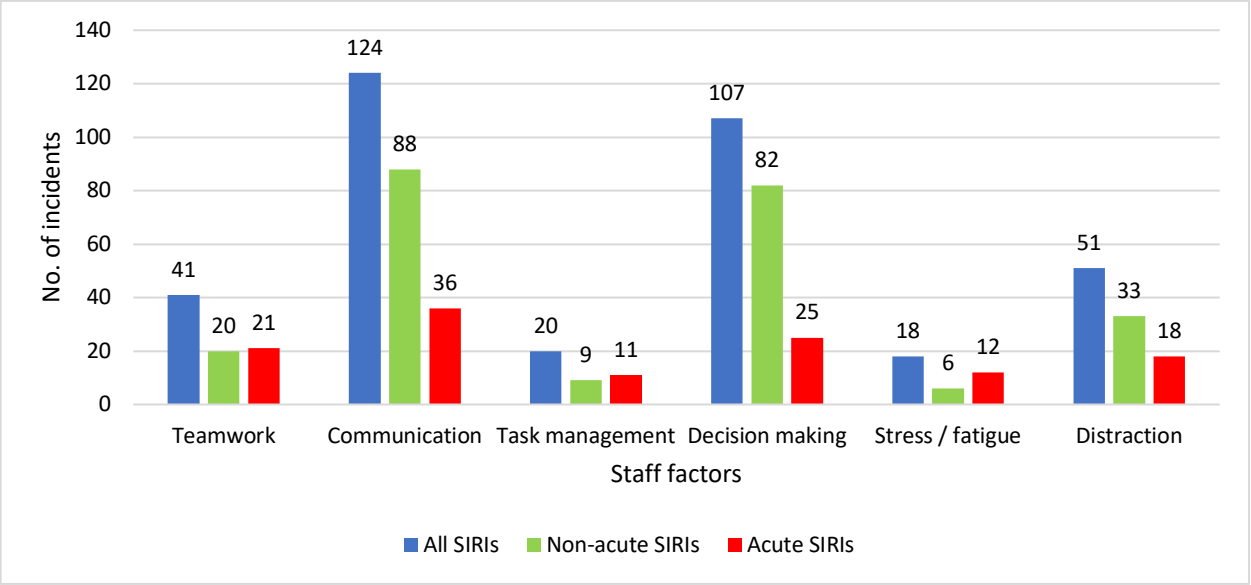


Figure 4-7: Staff factors identified as contributory in 167 SIRIs and including NTS, stress or fatigue and distraction.

4.7 WORK / ENVIRONMENT OR ORGANISATIONAL FACTORS

Staffing levels were recorded as a contributory factor where they were specifically highlighted in the incident reports (e.g. in many cases of pressure ulceration or falls reduced nurse staffing levels were cited as a problem). There were four attributes in this category and the incidence of each is shown in Figure 4-8.

Equipment and technology issues were contributory factors in 48 (39%) of non-acute SIRIs and included: inadequate equipment for the visualisation of the lumbar spine in a wrong level operation (case index no. 1); failure to use the SafeTx system in a wrong transfusion case because it was falsely believed to be faulty (case index no. 4) and problems with use of a foetal

scalp electrode and a CTG (cardiotocograph) machine in the maternity unit leading to inadequate monitoring of a hypoxic baby (case index no. 41). The electronic patient record was considered separately from equipment and technology because the OUHT is one of the few in the NHS at the moment with a ubiquitous, electronic system for the collection of patient data. There were 39 (23%) instances of problems with EPR across all SIRIs. This was more evident in the non-acute SIRIs than the acute SIRIs (18% vs 9% respectively) and included training issues and mixed paper and electronic documentation systems causing confusion in handover of care etc.

Errors in the use of standard operating procedures (SOPs) occurred in 149 incidents (89%) and were considered as one group but were subcategorized into: no SOP (10 [8%] in the non-acute and 6 [14%] in the acute SIRIs); SOP not used adequately or at all (91 [74%] in the non-acute and 29 [66%] in the acute SIRIs) or SOP inadequate / difficult to find (10 [8%] in the non-acute and 3 [7%] in the acute SIRIs).

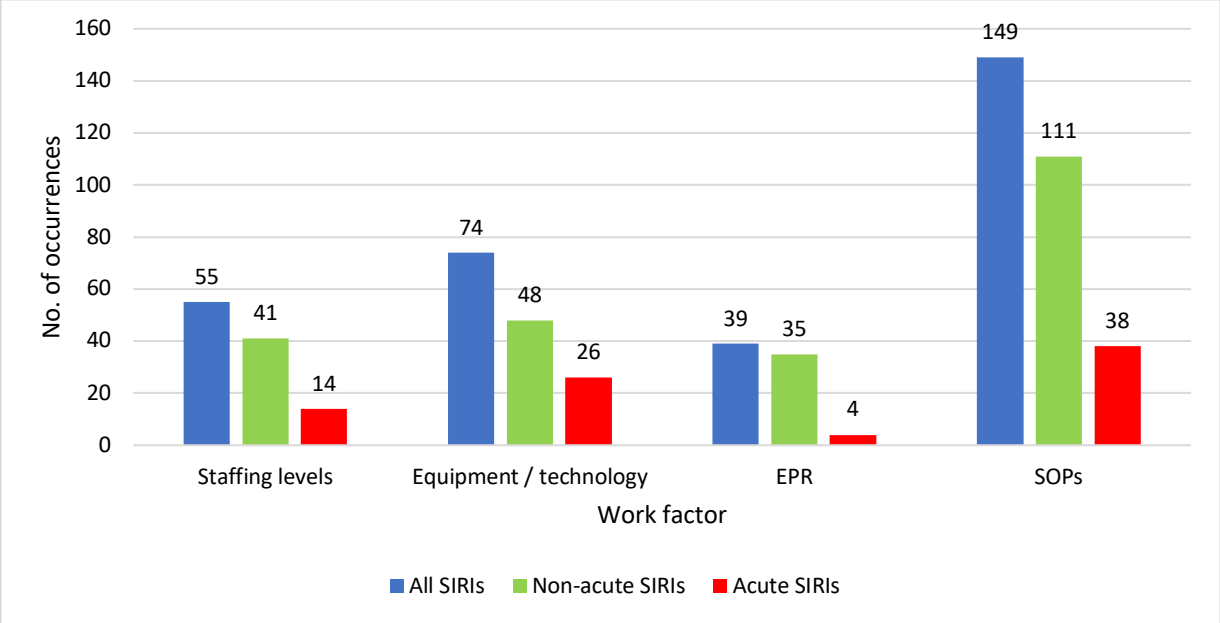


Figure 4-8: Incidence of work or organisational factors in 167 SIRIs (the OUHT's EPR system is considered separately from other issues with technology); EPR – electronic patient record, SOPs – standard operating procedures

4.8 GENERIC ERROR MODELLING SYSTEM (GEMS) FOR ACUTE AND NON-ACUTE SIRIS

Review of 167 SIRIs found six cases in which there was no error. Five were non-acute SIRIs which included: one fall where all appropriate measures were in place (case index no. 108); four incidents (case index nos. 138, 140, 159 and 162) where pressure ulcers developed in complex patients who were very unwell and were not deemed preventable and one of the acute SIRIs (case index no. 9) which was a case of insertion of a percutaneous gastrostomy tube for feeding in which a rare but recognised complication occurred and was promptly recognised and treated.

Error chains were produced for each SIRI and two examples are displayed below (Figure 4-9 and Figure 4-10). The first was an acute SIRI (case index no. 6) in which a patient admitted with abdominal pain had a delayed diagnosis of ectopic pregnancy resulting in an emergency, out of hours transfer to theatres on a different site in the hospital because she was decompensating secondary to extensive blood loss. The surgical team assembled for the WHO checklist but the consultant had not yet arrived and so missed this part of the safety check. The laparoscopy was started (with the consultant present) but, because he had missed an opportunity to discuss available equipment at the WHO checklist, he was unaware that the cautery machine was different from that in gynaecology. When he came to use it, he couldn't make it work and so assumed it was broken. The patient was deteriorating rapidly and so he performed an open procedure to stop the bleeding. This was a KB error, in the high stress moment the more effortful cognitive processes which would have allowed the surgeon to realise he *could* operate the cautery machine were bypassed by a decision to go to an open procedure (because that is a more straightforward option). The patient recovered well but was exposed to a higher risk

procedure with potentially more complications. The final error point is marked with a red arrow and the time from the error to the decision to switch procedures was approximately 5 minutes.

The second case was a non-acute SIRI (case index no. 8) was a patient was admitted with sepsis and a past history of dementia who suffered a fall on a geratology ward. He had been appropriately assessed as high risk using the Trust's falls assessment tool but inappropriate measures were used to safeguard him. He was placed in a side room (which made it difficult to observe him) and on a bed which was too high and had the bed rails elevated (putting him at risk of getting trapped). After the fall, which happened in the night, a busy trainee doctor was called to review the patient and found him difficult to assess because of his confusion. He made a suggestion that the patient's hip should be assessed more thoroughly in the morning but no X-Ray request was made. The next morning (day 2) this information was handed over to a different trainee doctor who ordered the X-Ray but did not follow it up and failed to escalate concerns to a more senior member of the team (there was no senior review of this patient). On day 3 of the patient's stay it was recognised that the X-Ray had still not been done and it was chased. Forty-six hours after the initial decision that an X-Ray was necessary a peri-prosthetic hip fracture was diagnosed. The final error point was taken from Day 2 when the message was handed over to the junior doctor on the day shift and was subsequently forgotten. This was an SB error (a lapse) – a simple task is forgotten by a trainee doctor who had many other distractions on the ward and the fracture diagnosis was substantially delayed.

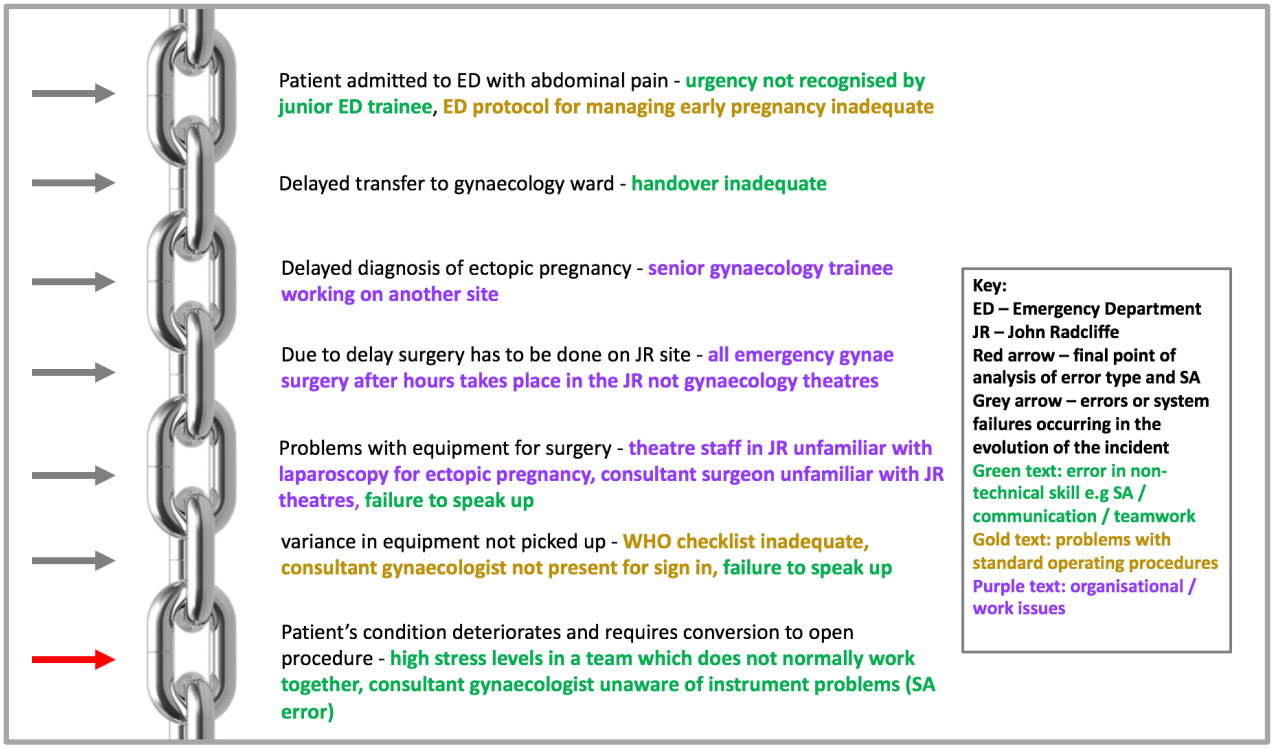


Figure 4-9: Error chain for SIRI in gynaecology (case index no. 6), grey arrows indicate errors or system failures occurring prior to the final key point of analysis which is highlighted by a red arrow

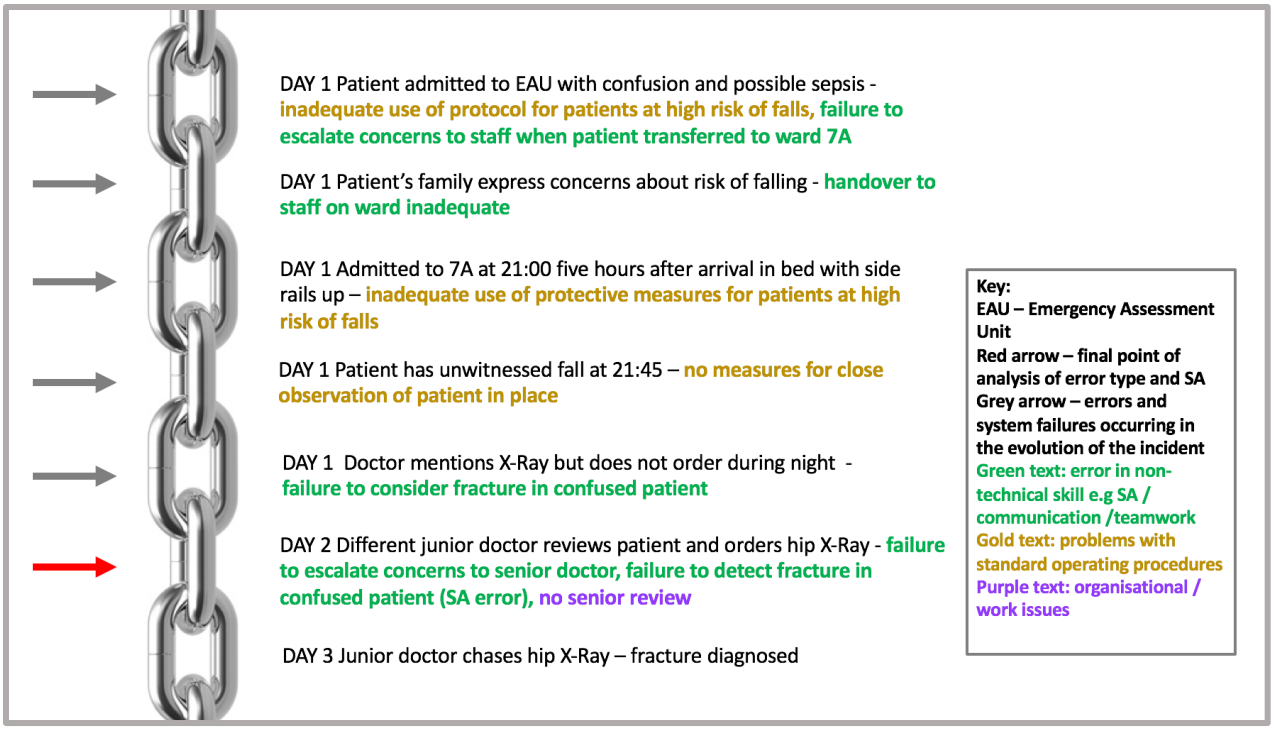


Figure 4-10: Error change for SIRI in geratology (case index no. 8); grey arrows indicate errors or system failures occurring prior to the final key point of analysis which is highlighted by a red arrow

Table 4-4 gives examples of error categorisation including the those in Figure 4-9 and Figure 4-10.

Type of Error	Description of Error (case index no.)	Evidence from SIRI report
Skill based slips and lapses **"errors which result from some failure in the execution (slip) and / or storage (lapse) stage of an action"**[21]	Lapse: patient falls on ward and hip fracture is undetected for 2 days (8 – see error chain below)	Initial assessment at night by trainee doctor who suggests hip x-ray but the message is not passed on clearly and the X-ray is forgotten
	Lapse: patient prescribed a time critical treatment for pulmonary embolism (Hospital Acquired Thrombosis) which is not given (25)	Drug prescribed out of hours but nurse unable to give as drug is not a stock item, forgot to inform doctor and doctor forgot to check
	Lapse: category 3 pressure ulcers develop on both heels due to lack of skin inspection (27)	Patient with dementia on a very busy ward, staff forget to undertake regular inspection of pressure areas
	Slip: during an emergency procedure at night a surgeon performs burr holes on wrong side of a patient's head (66)	Despite performing WHO checklist and confirming correct side for procedure a surgeon who has done extra shifts and has had a busy day starts the procedure on the wrong side of the head
	Slip: doctor sites a nerve block on the wrong leg (147)	At the point of siting the block there is a distraction in the anaesthetic room and the block is sited on the wrong side.
Rule-based mistakes **"the mistake arises from the application of a 'bad' rule or the misapplication of a 'good' rule [a rule of proven worth]"**[21] **Violations** **"deliberate deviations from those practices deemed necessary to maintain the safe operation of a potentially hazardous system"**[21]	A swab is left in a wound after completion of surgery (65)	An image intensifier was used to identify the site of the swab when a plain X-ray should have been taken for greater clarity – swab was not found until a plain X-ray was taken in recovery. (Application of a "bad" rule). Staff did not use swab accounting procedures correctly (misapplication of a "good" rule)
	Patient transferred post-procedure without being checked first and without adequate oxygen and appropriately skilled staff (119)	Post-procedure complication is undetected because rules around transferring patients are not followed (misapplication of a "good" rule)
	Patient transferred from another hospital for plasma exchange has a femoral line for the procedure and develops MRSA septicaemia (40)	SOP for caring for lines not followed – considered complex and time consuming to record visual inspection on electronic system (rule violated)

Knowledge-based mistakes **"mistakes arising from "the more laborious mode of making inferences from knowledge based mental models of the problem space"**[21]	Patient losing blood and deteriorating rapidly on the operating table, surgeon unable to use laparoscopic equipment available (in unfamiliar theatre) decides to default to open procedure. (6 – see error chain below)	Surgeon absent at WHO sign in and unaware that equipment available is not what he is used to, when the pressure mounts he is unable to perform the procedure using the laparoscopic equipment available and converts to an open procedure
	Patient post endoscopy is referred back to the surgical team by his GP with epigastric pain and admitted for investigation. The patient deteriorated and the actual diagnosis of myocardial infarction was delayed (152)	High pressure situation when patient deteriorates and despite an ECG being done the diagnosis was missed by the surgical team who were focused on an abdominal cause of the problem.

Table 4-4: Examples of error categorisation in 167 SIRIs using Reason's GEMS[21] model; case numbers in green are non-acute SIRIs; case numbers in red are acute SIRIs

Error type	Non-Acute SIRIs (%)	Acute SIRIs (%)
Skills based slips	9 (7)	9 (20)
Skills based lapses	56 (46)	6 (14)
Rule-based mistakes	49 (40)	11 (25)
Knowledge-based mistakes	4 (3)	17 (39)

Table 4-5: Error types (as classified by Reason's GEMS model[21]) in acute and non-acute SIRIs (no error ascribed in 1 acute and 5 non-acute SIRIs); knowledge based mistakes were significantly more common in acute SIRIs (see table 4-6)

Table 4-5 shows the incidence of slips lapses, RB and KB errors in acute and non- acute SIRI. Odds ratios were calculated to determine the likelihood of each type of error occurring in an acute incident and for the purposes of this analysis SB slips or lapses are combined into one group (see Table 4-6).

Error type	Odds ratio (95% C.I.)
Skills based (slips or lapses)	0.46 (0.23, 0.95)
Rules based	0.5 (0.23, 1.09)
Knowledge based	18.73 (5.83, 60.15)

Table 4-6 Odds ratios (with 95% confidence intervals) for the likelihood of different error types occurring in acute vs. non-acute SIRIs; calculations based on totals shown in table 4-5; knowledge-based mistakes were significantly more common in acute SIRIs.

Finally, 16 of the 21 KB errors (76%) in all the SIRIs and 13 of the 17 in acute SIRIs (76%) happened out of hours when reduced resources are available.

4.9 ANALYSIS OF SA ERRORS

Twenty cases were randomised for independent review in the initial phase of this study (described in the previous chapter) by both investigators and interrater reliability was calculated for all three levels of SA using Intraclass Correlation (SPSS statistics IBM™ V24.0). The results showed that at level 1 scores for both investigators were in absolute agreement, at level 2 the ICC was 0.78 and at level 3 it was 0.77 (an ICC value >0.70 constitutes good agreement[141]). The original plan had been for the case load to be split between the two investigators but it was not possible for the Deputy Medical Director to undertake the planned review for operational reasons and the complete analysis was undertaken by the author alone in the knowledge that our agreement was good. Where any uncertainty existed in specific cases a joint review was undertaken and consensus reached.

Analysis of SA at all three levels (including sub levels for each) was completed for 167 SIRIs. SA errors were detected in 96% (161/167). There were 6 incidents where no loss of SA occurred and these were the same incidents in which no errors were ascribed (case index nos. 9, 108, 138, 140, 159 and 162 as described above). For the remaining 161 cases SA errors were present at either level 1,2 or 3 (see Table 4-7).

Examples of SA error categorisation in acute and non-acute SIRIs are included in Table 4-8.

SA Level	Incidence in 123 non-acute SIRIs No. (%)	Incidence in 44 acute SIRIs No. (%)
1: Failure to perceive information	**92 (75)**	**22 (50)**
Data not available	3 (2)	0
Data hard to discriminate or detect	14 (11)	7 (16)
Failure to monitor or observe data	65 (53)	7 (16)
Misperception of data	6 (5)	4 (10)
Memory loss	4 (3)	4 (10)
2: Improper integration or comprehension of information	**22 (18)**	**15 (34)**
Lack of / incomplete mental model	15 (12)	8 (18)
Use of incorrect mental model	7 (6)	7 (16)
Over reliance on default values	0	0
3: Incorrect projection of future actions of the system	**2 (2)**	**6 (14)**
Lack of or incomplete mental model	2 (2)	6 (14)
Over-projection of current trends	0	0

Table 4-7 SA errors in 123 non-acute and 44 acute SIRIs; numbers in bold represent totals for each level of SA with sub-levels displayed below.

SA level	Non-acute SIRIs (case index no.)	Acute SIRIs (case index no.)
1	Data were filed incorrectly in an abdominal aortic aneurysm screening database so that clinicians could not see data on incidental findings and patients were not followed up appropriately – level 1 error: data were not available (32) A 15 yr old patient with cerebral palsy (with learning difficulties and speech impairment) developed pressure ulcers on both heels after surgery to both legs with full length plaster casts in place – level 1 error: data were hard to discriminate or detect (112) An 83 year old patient was admitted to a ward with haematemesis and incorrectly assessed as low falls risk, inadequate measures were put in place to observe the patient and she subsequently fell and fractured her hip – level 1 error: data were not observed or monitored (57) 19 patients in a pre-assessment clinic had to return the following day for repeat samples to be taken as the phlebotomist had misread and scanned the labels on the patients' notes rather than scanning the patients' wristbands as per protocol – level 1 error: misperception of data (126) A 72 year old patient was sent home without the medication he needed to reduce gastric acidity and was readmitted with a catastrophic haemorrhage, the doctor prescribing the drugs to take home knew the patient but was not the doctor who wrote the discharge summary and on a busy day forgot that he should be on a proton pump inhibitor – level 1 error: memory loss (127)	A 90 year old patient was having an operation on her hip and a femoral block was sited on the incorrect side, the mark on her leg indicating the side had been covered up – level 1 error: data were hard to discriminate or detect (144) The incorrect radiopharmaceutical was injected into a patient for a scan, no final check of the label was undertaken – level 1 error: data were not observed or monitored (101) A tray of unsterile instruments was used for an organ retrieval, colour changing tape misperceived by scrub nurse, indicator tape on instrument trays goes from pink to brown after autoclave but can degrade and look browner (i.e. sterile) – level 1 error: misperception of data (83) Patient was recovering on ICU but remained agitated and repeatedly disrupted oximeter reading by touching the probe, nurse removed pulse oximeter whilst she was sitting with him but forgot to hand on that data when she went for a break and the patient lost consciousness – level 1 error: memory loss (34) (N.B. no acute SIRIs had a level 1 SA error where data were not available)
2	A 16 year old was seen in ED with severe abdominal pain after a football injury and was mistakenly allowed to leave the department whilst the pain was still ongoing but incomplete and confusing pain assessments hampered decision making, he required emergency surgery the following day for a perforated colon, incorrect decision made to discharge on basis of incomplete understanding of situation – level 2 error: incomplete mental model (155)	A patient admitted with abdominal pain had a delayed diagnosis of ruptured ectopic pregnancy which led to emergency surgery in an unfamiliar theatre. Consultant surgeon had incomplete mental model (thought laparoscopic equipment would be what he was used to) and found he could not work out how to use the cautery device and therefore performed a laparotomy – level 2 error: incomplete mental model (6)

	A 69 year old patient was admitted after a fall and spinal fracture on top of pre-existing cervical stenosis (which was causing diaphragmatic weakness). He was prescribed opiates for pain (he refused surgery) and had a respiratory arrest. His oxygen requirements were recorded as rising but it was incorrectly assumed that this was secondary to his diaphragmatic weakness and not the opiates – level 2 error: incorrect mental model (17)	A patient was admitted for observation on the midwifery unit, midwife observing her did not escalate foetal observations when established labour began as did not think the patient had transitioned (incorrect mental model), delayed transfer to theatre with placental abruption – level 2 error: incorrect mental model (22)
3	An 87 year old patient with diabetes and renal impairment was admitted to an ENT ward with a severe salivary gland infection, her diabetes was difficult to manage, her renal function worsened and her potassium levels rose dangerously. Management of the high potassium was not aggressive enough because there was a failure to understand the gravity of the situation and the speed with which things could worsen she had a cardiac arrest – level 3 error: inadequate mental model (31)	Foetal distress was detected in a patient on delivery suite, delayed recognition of severity of the situation by trainee obstetrician who recognised that a caesarean section was necessary but felt it was a category 2 (less urgent situation) and failed to project what the delay would mean to the foetus - level 3 error: inadequate mental model (41)

Table 4-8 Examples of SA categorisation at all three levels and relevant sub-levels (see table 4-7) for non-acute and acute SIRIs.

Odds ratios for the likelihood of SA errors at each of the three levels in acute care SIRIs were calculated and are shown in Table 4-9.

SA Level	Odds ratio (95% C.I.)
1	0.34 (0.16, 0.69)
2	2.37 (1.09, 5.16)
3	9.55 (1.85, 49.31)

Table 4-9 Odds ratios (with 95% confidence intervals) for the likelihood of SA errors at three levels in acute vs. non-acute SIRIs; calculations were based on totals at each level as shown in table 4-7 (i.e. sub-levels were not used because numbers were too small)

4.10 DISCUSSION

This study has taken a cohort of serious incidents which occurred over one year in a large acute care Trust in the NHS and produced a comprehensive analysis of the attributes including clinical

setting, timing and outcome as well as considering the incidence of error by type (SB, RB or KB) and by SA errors.

This discussion section is divided into three sections which consider:

- Incident attributes, including where and when, patient, staff and work factors
- Error type
- SA error

4.10.1 INCIDENT ATTRIBUTES

4.10.1.1 SITE OF INCIDENT

Most (73%) of the SIRIs occurring in ward settings evolved over longer time periods (e.g. development of pressure ulcers and hospital acquired thrombosis) and in outpatient settings (e.g. diagnostic delays of several years). The context of these incidents is one in which decisions are less commonly time critical and where there are opportunities to reflect and consider options as well as ask for advice.

The acute SIRIs were more common in theatre, ICU, ED and maternity than ward, outpatient or radiology settings (61% vs 39% respectively). These are clinical areas in which patient turnover is more rapid and clinical acuity is higher.

4.10.1.2 TIME OF DAY

The results showed that 19% of the SIRIs happened out of hours but that an acute SIRI was six times more likely to happen out of hours than a non-acute SIRI. This finding resonates with

recent data from NCEPOD[189] along with research from the NHS revealing higher morbidity and mortality out of hours than in hours[190–193].

4.10.1.3 TIME FRAME FOR EVOLUTION OF INCIDENT

Analysis of the time frame over which the incidents evolved revealed a subset of SIRIs that happened over a shorter time frame of less than 8hrs (which is representative of a normal shift length for junior doctors in the NHS) and 31 of these (70%) evolved in less than an hour. This subset of acute SIRIs possessed key characteristics in terms of:

- demographics of the incidents (where and when they occurred) i.e. mainly in high turnover settings with high levels of clinical acuity such as theatre and ICU and out of hours (see above)
- error type and SA errors (these are explored below)

4.10.2 PATIENT FACTORS

The most obvious difference in patient factors between the acute and non-acute SIRIs was the clinical acuity of the case. Errors are more common in emergency settings[194,195] and in elderly patients with more comorbidities.[194,196,197] 75% of acute SIRIs were categorised as urgent or emergency as opposed to 28% of the non-acute SIRIs and the urgency of a situation has a detrimental impact on decision making and team performance.[34,198,199] There were also obvious differences between acute and non-acute SIRIs for presence of multiple comorbidities. However, a more detailed examination of these cases revealed that although there were more non-acute SIRIs in which patient co-morbidities contributed to the evolution of the incident (47% vs 20%), 93% of them occurred in patients who developed pressure ulcers, or had falls or HATs. In these case the co-morbidities commonly led to challenges in providing care (e.g. a

patient with dementia who fell despite precautions being in place or a patient with morbid obesity where it was difficult to reposition regularly to prevent pressure ulceration occurring). Conversely in the acute SIRIs, in all nine cases where patient comorbidities contributed the team's decision making was impacted by the additional complexity. For example:

- Case index no. 157 - a 47 year old patient with learning difficulties was admitted for control of his diabetes and the fall in his blood pressure and tachycardia were put down to sepsis secondary to high blood sugar when he had actually had a myocardial infarction which was evident on his ECG
- Case index no. 156 - a patient with congenital airway problems and a complex cardiac history had a valve replacement. Three days later he had a cardiac arrest. The abnormal airway anatomy made ventilation difficult and diverted attention from the presence on a CT scan of a mediastinal haematoma and meant that a sternotomy was not performed when it should have been

These challenges in decision making either due to clinical acuity or case complexity are explored further in the following sections.

4.10.3 STAFF (NTS, INDIVIDUAL OR TEAM) FACTORS

4.10.3.1 IMPACT OF NTS ON SIRIS

Errors associated with suboptimal teamwork were common in this study in keeping with findings in other studies.[169,200–203] The issues with teamwork were, however, always associated with poor communication. Communication has been implicated in many errors in healthcare and commonly fails at interfaces between multidisciplinary teams[204–207] and good communication plays a vital role in raising team situation awareness.[208] This analysis of SIRIs

also found that MDTs were more commonly involved in acute than non-acute SIRIs (68% vs 43%).

4.10.3.2 STRESS, FATIGUE AND DISTRACTION

Stress and fatigue were more common in acute SIRIs than non-acute SIRIs (27% vs 5%) but the mention of stress as a contributory factor was low in the reports and it has been shown that healthcare professionals downplay the impact of stress and fatigue on performance.[209] Distraction was also more evident as a contributory factor in acute SIRIs (41% vs 27%) with rapidly changing circumstances forcing abrupt changes of focus and complicating the management of acutely unwell patients.[210,211]

4.10.4 ERROR TYPE

Since the publication of To Err is Human,[16] multiple tools for the categorisation and analysis of medical error have been developed but most of them have focused on the context and clinical category of the error [8,11,194,212] rather than on the cognitive processes that may underlie them. Human factors based tools for the analysis of errors have been developed in military and civil aviation[213] chemical plant[214] and nuclear power settings[215] in advance of healthcare but more recently there has been research in cognitive errors in primary care,[170] in medical trainees,[35,187,216] in anaesthesia[217] and hospital settings.[218]

This study has used Reason's GEMS system[21] to classify errors in an acute care hospital and found multiple errors in each of the incidents from each of the three categories (SB, RB or KB mistakes).

SB errors are deviations from intended actions due to execution (slip) or storage (lapse) failures whereas RB and KB mistakes occur at the level of intention formation.[219] Errors at the level of SB and RB processing are more common simply because this type of processing is much more common in human task performance. Reason highlights an important point, however, about the ratio of error to opportunity in SB, RB and KB levels: "skill-based and rule-based processing are the hallmarks of expertise…. When expressed as proportions of the total number of opportunities for error at each performance level, the percentage of errors in the SB and RB modes will be very much smaller than at the KB level of processing."[219] Furthermore, the likelihood of making an effective correction is greater at SB than KB levels.[220]

The proportion of errors at SB, RB and KB levels in this study is similar to those in a study of serious incidents in intensive care[221] and in junior doctors in simulated scenarios.[222]

SB errors occurred more commonly in the non-acute SIRIs and the majority were lapses (46%). Examples included forgetting to undertake regular pressure area checks, forgetting to follow up on advice regarding anomalous data on a scan and forgetting to perform VTE assessments. Examples of slips included assigning the incorrect consultant to a patient resulting in delayed referral to the appropriate team and a letter for one patient being put in the notes of another. These are not the type of error which would be amenable to experiential training (including simulation) and in many, the contributory factors highlighted above (such as staffing problems and stress or fatigue) were more important in the evolution of the incident.

RB mistakes can be further broken down into three broad categories: misapplication of good rules, application of bad rules and rule violations. RB mistakes were more common in non-acute than acute SIRIs (40% vs 25% respectively). In 89% of all the SIRIs problems were found with standard operating procedures (SOPs) - not all of these were evident at the final error point but

many were implicated at an earlier stage in the evolution of the incident (see below). Not all rules, however, are explicit SOPs – they can be stored schemata which are used to compare with current situations (the "I've seen this before or done this before" feeling) and examples of this included: a bad rule for instrument change in a robotic surgical operation (case index no. 18) and the use of piece of x-ray equipment known to be inadequate for the purpose (because "it's all we have", case index no. 1).

Violations have been defined as "deliberate (but not necessarily reprehensible) deviations from those practices deemed necessary to maintain the safe operation of a potentially hazardous system"[223] and only 5 were found in this cohort of cases.

KB mistakes were found to be 19 times more likely in acute SIRIs. These mistakes involve cognitive effort and this becomes more challenging under conditions of time pressure. Dörner has described an "intellectual emergency reaction"[224] in critical situations (like that experienced by the surgeon in the gynaecology case [index no. 6] above) whereby reflexive rather than intellectual activity predominates and decisions are made without pause for reflection. The mode and median times from final error point to incident occurrence in the knowledge based mistakes was 5 and 119 minutes respectively. These types of error are amenable to experiential training interventions which allow repeated practice to develop automaticity of action and a reduction in cognitive load.[113,114]

4.10.5 WORK / ENVIRONMENT OR ORGANISATIONAL FACTORS

Inadequate staffing levels have been associated with low morale, poor staff and patient satisfaction and poor patient care.[225,226] Problems with staffing levels were cited as contributory factors in one third of all SIRIS (this proportion was the same in acute and non-acute SIRIs).

Human interaction with mechanical or electronic devices has been intensively studied in other dynamic workplace settings and user-centred design plays a vital role in the creation of safe systems.[155,227–229] It is unfortunately the case that design to improve human performance is not much in evidence in today's NHS. In a recent (unpublished) audit we found 37 different monitor interfaces across the four hospital sites in the OUHT (only the ECG trace was standardised to a colour [green] and position on the screen [at the top]). The majority of problems with the OUHT's EPR system related to the inconsistent use of electronic recording across clinical areas and the persistence of paper systems (often in conjunction with EPR).

Checklists and SOPs have been used for decades in military and civil aviation (e.g. pre-take-off checklists) and are now more common in healthcare, especially surgery[230] and, more recently, all invasive procedures.[231] However, they will not improve safety if they are not used properly or at all.[232–234] Overall there were 149 (89%) incidents in which SOPs were implicated. The majority of these issues stemmed from misuse or no use of the SOP (120 incidents [81%]) and in the remainder they were not present or not fit for purpose

4.10.6 SA ERRORS

Analysis of this cohort of incidents revealed a higher rate of SA error (96%) than other similar studies in healthcare which considered SA errors in voluntarily reported anaesthesia incidents[78] (SA errors were identified in 81%) and in closed anaesthesia malpractice claims for death and

brain damage[79] (SA errors were identified in 74%). However, these were a subgroup of incidents reported in the OUHT which have been investigated because they were highlighted as particularly serious (see above) and it was not, therefore, surprising that multiple SA errors across sites, professional groups and times were found.

The studies highlighted above only considered SA as a contributory factor and whilst it is empirically correct to say that SA errors were present, it is a futile exercise in isolation if we wish to understand the broader context of SA and error in the clinical workplace. This study has considered patient, staff and work related factors as well as error type and SA in order to do this. For example, in the majority of the cases of pressure ulcers (non-acute SIRIs) the demands of work, existence of poorly designed and utilised standard operating procedures (SOPs) and high levels of junior staff with little experience of the management of pressure areas were more important features of incident causation, although loss of SA was apparent in all but four of them (as explained above).

SA errors occurred in 96% of cases in this study but errors at the levels where greater cognitive effort was required (i.e. level 2 (comprehension) and level 3 (projection)) were more likely to occur in the acute SIRIs (twice and 9.5 times respectively). Furthermore in 18 of the 21 KB mistakes level 2 and 3 SA errors were detected and of the 8 level 3 SA errors 6 occurred in this KB group and all of those were acute SIRIs. It is in these more time-pressured incidents that a focused experiential training intervention involving simulation may improve individual and team SA and performance.[31,168]

4.11 STUDY LIMITATIONS

The OUHT has been required to submit patient safety incidents since 2003 when the NRLS came into operation in the NHS (all hospitals are required to report patient safety incidents and the NRLS is now the largest healthcare reporting system in the world[235]). The OUHT's reporting rates are shown in the previous chapter and the Trust is in the top half of the group. Whilst it is tempting to suggest that this reflects a better than average reporting culture a recent review has advised caution in interpreting these data as no particular hospital characteristics were found to be significantly associated with overall reporting rate.[235] Underreporting of incidents is also a key concern in any voluntary reporting system (such as Datix) and has been highlighted in healthcare[236–238] and presents a clear limitation for retrospective analysis of incidents.

The SIRI reports reviewed in this study were presented in the standardised NPSA format (see previous chapter) and this made the process of analysis more straightforward. It was, however, evident that a human factors approach was not consistently applied. An analysis of the quality of the reports is outside the remit of this thesis but similar inconsistencies have been observed in other studies in healthcare and in different domains such as aviation and finance.[239–241] Indeed in one review of aviation incidents the authors were able to search their incident reports using SA as a search term[108] but it did not appear once in these reports. Whilst the NPSA guidelines have been in existence since 2010 there is evidence that the quality of incident reporting is variable within the NHS.[242]

Studies on error in healthcare have often used retrospective case review methodologies[8,11,194] and the two more recent studies on SA in healthcare[78,79] included above have used similar techniques. This study has used a mixed methods retrospective review of incident reports with the qualitative thematic analysis of the 167 incidents undertaken mainly by the author alone.

Author reflexivity is an important consideration in any research and no researcher can be absolutely objective.[179] In the context of this work possible sources of bias exist for both researchers and included both our clinical experience in cardiology and anaesthesia which may have biased decisions on specialties outside our own (although the extensive hospital-wide interdependencies of both cardiology and anaesthesia have led to exposure to all the specialties represented in the cases we reviewed for this study). This is mitigated by the fact that all the incident reports were led or supported by staff with context specific expertise and both my own and the deputy medical director's experience in investigating medical error in the OUH and other NHS hospitals.

Identifying risk factors for critical incidents forms a vital first step in understanding how to prevent them and defining incident attributes using thematic analysis was the method chosen for this study. Decisions made for each of the incident attributes as well as error types and SA error could also be subject to bias but attempts to limit this were made by using clear definitions (as described above, many of which have been used in other studies of error in healthcare) and clarifying any ambiguity with joint discussion or referral to context specific experts.

4.12 CONCLUSIONS AND FUTURE DIRECTIONS

This study has shown that situation awareness errors occurred in 96% of incidents in an acute care Trust in the NHS and that errors in the more effortful cognitive pathways (level 2 and 3 SA errors and RB and KB mistakes) are more common in incidents which evolve over shorter time frames. SA errors played a key role in the majority of rapidly evolving incidents but, whilst present, they were less important in those that took days or weeks, where staffing levels and problems with standard operating procedures were more significant.

The overarching purpose of this thesis is to develop insights into the role of reduced SA in healthcare, particularly in the dynamic, often stressful settings the author works in and to consider how we might design training interventions to improve SA and support healthcare professionals in managing critical situations which evolve over much shorter timeframes. There is some evidence from aviation and anaesthesia to suggest that focused SA training is a promising solution for improving individual and team SA particularly in high pressure settings such as combat training for military fighter pilots.[168] In order to determine the successful outcome of a training intervention, however, valid and reliable measurement tools must be used. Measurement of SA in healthcare is most commonly undertaken using tools designed for the assessment of several categories of NTS including SA (e.g. teamwork, decision making or communication). A review of these tools and their strengths and limitations will form the subject of the following chapter.

CHAPTER 5 SYSTEMATIC REVIEW OF THE VALIDITY, RELIABILITY AND USABILITY OF TOOLS FOR NON-TECHNICAL SKILLS ASSESSMENT IN SIMULATED OR REAL CLINICAL ENVIRONMENTS IN HEALTHCARE

5.1 INTRODUCTION

Evidence that errors in NTS are common in adverse incidents in healthcare has been accruing over the past two decades and has been discussed in depth in Chapter 1.[8,11,12,194,243]

SA is most commonly measured in healthcare using observer-based indirect techniques incorporating SA as one category amongst several other NTS. Interest in evaluating and enhancing NTS in multi-professional teams of healthcare workers has been increasing in line with concerns highlighted in studies of error in healthcare. A number of tools are now available for measuring them with many of the early examples adapted from the civil aviation field.[110,145,152,244] Concerns about the measurement properties of these tools (including their validity and reliability) have been raised by educational and research communities.[114,160,161,245] Assessment of healthcare professionals, particularly in high stakes settings such as examinations or interviews, requires rigorous attention to the quality of the tool being used to make that assessment if it is to be objective and fair. Furthermore, the choice of an appropriate tool for NTS assessment may be hampered by the large number available for different settings in healthcare.

This systematic review of the NTS assessment tools for healthcare seeks to provide a clearer understanding of the range, purpose, validity and measurement properties of published tools.

5.2 OBJECTIVES

The objectives were:

1) To provide an overview of assessment tools for measurement of NTS in healthcare professionals or students in simulated or clinical environments
2) To evaluate the methods used in developing the tools
3) To analyse the evidence for the validity and reliability of the tools
4) To evaluate ease of use, and training required for each of the assessment tools
5) To provide guidance in selecting a tool for a given context

5.3 METHODS

This systematic review was registered with Prospero (ref. no: CRD42017055445). Peer-reviewed studies were identified by search of the electronic bibliographic databases: PubMed, Embase, CINAHL, ERIC, PsycNet, Scopus, Google Scholar and Web of Science. A search of the grey literature was made via OpenGrey, ProQuest, AHRQ, the King's Fund and the Health Foundation. A manual search of the reference list of identified relevant articles was also conducted. No further searches were conducted after March 2017.

All reviewed articles were assessed using criteria defined by Hawker et al for mixed qualitative and quantitative research studies.[246] The assessment questionnaire and a detailed search strategy are included as Appendices 7 and 8.

5.4 SYNTHESIS OF RESULTS

Papers with potential for inclusion in the review on the initial search were first screened for relevance, by review of the title, and then by abstract review (see Figure 5-1. for the PRISMA review process). Papers with a relevant title and abstract were retained for full review. Papers without any assessment of validity or reliability for the NTS tool being used were discarded. Where papers were not retained for review, their reason for non-inclusion was recorded.

The first stage of the screening process was conducted for all papers in pairs (HH and PG; HH and JR or PG and JR) – where any disagreement was encountered a decision was made by the reviewer who was not a member of the original pair. Full text articles were acquired for all abstracts put forward for further analysis. These were divided between the three reviewers for initial assessment and any ambiguities arising regarding inclusion were discussed and agreed together. The final in depth analysis was then undertaken by HH and PG with JR acting as final arbiter. All first authors were contacted by email, on two separate occasions, to seek additional unpublished information.

Most of the tools had already been given a name (for example TEAM – Team Emergency Assessment Measure[154]) and, if not, we devised a name based on an approximation of the purpose of the tool (e.g. Anaesthetic trainee NTS[247]). A list of acronyms for all the tools in this review can be found in Appendix 9.

The NTS assessed by the tools were usually described in categories e.g. communication, teamwork and leadership etc. which were underpinned by behavioural markers (e.g. TEAM[154], OTAS[248], Oxford NOTECHS[249], NOTSS[250] and Ottawa GRS[251]) but some described an inventory of behaviours relevant to the context or professional group being analysed (e.g. UTBMNR[244],

MHPTS[252], TBR[253]). We classified NTS into the five most commonly occurring categories: communication; leadership and/or teamwork; situation awareness; decision making and task management. We also included an "other" section to capture elements not ascribable to one of these categories. Examples where additional behaviours were assessed included: professionalism;[254,255] "environment in the room"[256] and stress and distractors.[257] Where descriptors of behaviour were essentially a sub-category of one of the five domains they were included under the relevant heading e.g. cooperation was included under teamwork and vigilance under situation awareness.

Studies were analysed and scored over two broad domains: the method of development of the tool and the process of psychometric evaluation of the tool (including validity and reliability and any assessment of usability and training requirements). Where the *original* development and psychometric testing of a tool was described in more than one publication the data from all relevant papers were analysed, as long as at least one member of the original research team was involved.

Considerable variability was found in method of tool design and psychometric assessment in this study in line with previous systematic reviews of assessment.[245,258,259] Evidence of validity was classified (where possible) into domains described by the American Educational Research Association:[260] content (i.e. test items are representative of the construct of interest); response process (i.e. evidence of data integrity including: rater training, methods for scoring and data entry); internal structure (psychometric properties including: rater reliability and item correlations); relations to other variables (e.g. evidence that the test can discriminate between candidates or convergent evidence i.e. that the results are related those from a tool measuring another, similar construct). Cook et al[245] have highlighted the difficulty of applying

instruments used for clinical studies such as STARD[261] (Standards for Reporting Diagnostic accuracy) and GRRAS[262] (Guidelines for Reporting Reliability and Agreement Studies) in the context of assessing tools for educational assessment. To assist educators in selecting tools for NTS assessment a more pragmatic approach was adopted and tools have been categorised in terms of context of use, method of design (including method of scoring), psychometric testing and usability (see Table 5-1). The attributes assessed were developed iteratively by the authors and informed by: the initial study assessment questionnaire (see above and Appendix 8); our experience as clinicians and educators and guidance on design of educational assessment tools[263] (including validity and reliability[138,140,264,265] and team training assessments[266]). The scoring system was also tested for internal consistency and interrater reliability (see results below).

Method of tool design and context of use	0	1	2	3
Applicability	Non-healthcare	Subtask only	One specialty or discipline	Multiple specialty areas or disciplines
Environment	Non-healthcare	Simulation	Real clinical settings	Simulation and real
Range of NTS	Unstructured descriptors of NTS	2 structured / distinguishable categories	3 structured / distinguishable categories	4 or more structured / distinguishable categories
Subject matter experts involved	Not reported	Single clinical discipline involved	MDT / professional involvement	MDT / professional clinicians + HF or psychology expertise
Response process (evaluation of scoring system)	Not reported	Description of scoring system choices	Explanation and justification of scoring system choices	Justification of choices and comparison with other scoring systems
Psychometric testing /usability	**0**	**1**	**2**	**3**
Construct validity: content*	Not reported	1 item of evidence	2 items of evidence	3 items of evidence
Construct validity: relation with other variables†	Not reported	1 item of evidence	2 items of evidence	3 items of evidence
Reliability of tool§	Not reported	1 item of evidence	2 items of evidence	3 items of evidence
Usability of tool¶	Not reported	1 item of evidence	2 items of evidence	3 items of evidence

Table 5-1: Scoring system for method of design, psychometric testing and usability of NTS assessment tool. MDT= multidisciplinary team. HF = human factors. Validity is described as per standards from the American Educational Research Association[260] other terms for validity used by original authors are included here in parenthesis. *Evidence of content validity included: literature review of relevant NTS, use of a Delphi method with an expert panel and formal task analysis. †Evidence of construct validity for comparisons with other variables included: expert-novice comparison (discriminant validity), comparison with the same or a related construct (concurrent or convergent validity) and comparison with a future performance (predictive validity). § Evidence of reliability included: internal consistency, inter-rater reliability and intra-rater or test-retest reliability. ¶ Evidence of usability included: specification of training, quantitative assessment of usability and qualitative assessment of usability.

5.4.1 RISK OF BIAS

Data analysis and interpretation was undertaken with an awareness of the risk of bias.

Repeated reflection on potential sources of bias in the context of personal beliefs and values (researcher reflexivity[179]) was integral to the iterative review of the studies in this review.

Study selection bias was minimised through use of a systematic search method.

Potential bias for the authors in reviewing the assessment tools included:

- Familiarity bias: four of the authors are active educators in simulation based education, JR was the author of one of the tools (ANTS-AP[267]), CV has been involved in the development of tools for NTS assessment [268–270]). HH,PG and JR have been trained to use the ANTS assessment tool
- Availability heuristics: HH, PG and JR are practising anaesthetists, as such our training and clinical experience is largely in theatre and ICU settings
- Anchoring bias: the order in which the papers were reviewed and the organisation of information presented in each study may influence decisions made in assessing the tools

Mitigations for these risks included development of a scoring system for the tools as described above, review by more than one author and repeated re-examinations of the papers in random order.

5.5 RESULTS

The screening process is described in Figure 5-1 as per PRISMA guidance. All articles included for review were observational studies of healthcare professionals or students in simulated or real clinical settings.

Seventy-six unique tools for the assessment of NTS in healthcare were identified to be suitable for inclusion in the review. These were described in 116 papers. The first tool was developed by Gaba et al[145] in North America. Subsequently most tools have been developed in North America (35 tools), followed by Europe (31 tools), and Australasia (eight tools). One tool was developed in Colombia[271] and one in Israel.[272]

Most tools were developed de-novo, but some were explicitly based on tools developed by other groups (e.g. NOTSSdk[273] or OTAS-D[274]) and some relied on data gathered in the original tool (e.g. OTAS-S[271] or F-Team [275]). Self-assessment tools were excluded because, whilst they may be useful in formative settings, self-assessment of NTS is inaccurate and unsuitable for use in high stakes settings.[151]

Overall considerable variation was found in both methods of tool design (including NTS measured and purpose, context and environment of use) and evidence of validity reliability and usability.

The method of scoring used to provide a hierarchy of tools had good internal consistency (Cronbach's alpha: 0.72) and interrater reliability (weighted kappa: 0.95).

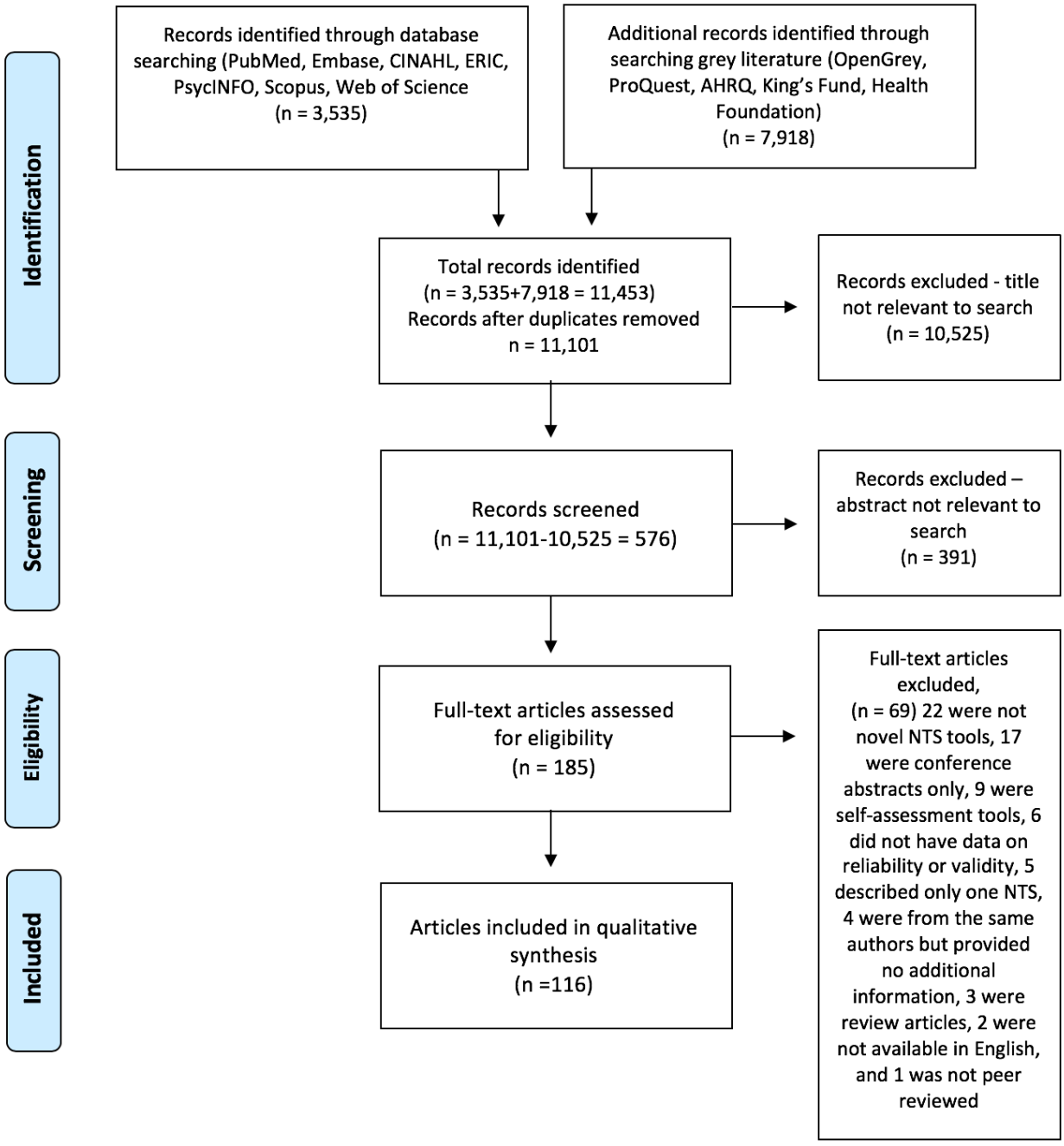


Figure 5-1: PRISMA diagram for systematic review of tools for the assessment of NTS assessment in healthcare

5.5.1 METHODS OF TOOL DESIGN AND CONTEXT OF USE

Methods of reporting observations varied. For example, number of observations made using the tool (e.g. BMS-NNTS[276] and EPOC[277] include an assessment of frequency of interactions), or number of participants or teams observed (some had large numbers of observations or

participants[28,277,278] and others fewer[274,279,280]) and some were individual or team assessments or both (see Table 5-2). Consequently, it was difficult to make meaningful inferences between the studies.

Most assessment tools (37 [49%]) had been designed for use with multidisciplinary teams; 27 (36%) were for single specialty postgraduate healthcare professionals; 8 (10%) were for the assessment of healthcare students and 4 (5%) were for multi-specialty postgraduate doctors (see Table 5-2 and Appendix 10).

The environments in which the tools were designed and tested varied but fell under two broad domains: simulated or real clinical settings.

NTS categories assessed were also variable. Communication was assessed in every tool although not always as an isolated category (e.g. Oxford NOTECHS[249] and ANTS[152]). Teamwork and leadership were the next most commonly included categories (74 [97%] of tools), situation awareness was assessed in 66 (87%), task management in 61 (80%) and decision making in 36 (47%).

Data for the 50 top scoring tools are shown in Table 5-2 with details for the remaining 26[145,254,272,274,275,277–279,281–302] provided in Appendix 10.

5.5.2 PSYCHOMETRIC TESTING AND USABILITY

The argument based approach to validity[138,139,303] was used to assess the tools but this was limited by the variability in the provision of evidence. All tools assessed content validity in some form and the next most common assessment was relation to experience or educational level of participants (47 tools [62%]). Tests of relationship with tools measuring similar, related constructs (25[33%]) were more common than those testing tools against others

measuring the same or related constructs (these tests were variably termed convergent or concurrent validity) (11 [14%]) and only three groups considered predictive validity in the sense of ability to predict future performance.[281,304,305]

Reliability was most commonly assessed with interrater testing (61 tools [80%]) or internal consistency (41 tools [54%]). Only 11 studies (14%) considered test-retest reliability.

Some authors went to great lengths to analyse usability and generated qualitative and quantitative data from questionnaires or interviews (which informed the development and deployment of their assessment tools).[152,256,257,267,306–312]

Recommendations for training were described in very different ways, from those who have designed bespoke courses for their tools (e.g. NOTSS, [153] OTAS,[313] and MINTS-DR[314]) to those where a tool was designed with a specific remit of not requiring much training to use it (TEAM[154] MHPTS[252], PETRA[308] and CTS[315]).

Table 5-2 shows data for method of development and evidence of validity, reliability and usability for the top 50 ranked tools. This table provides an overview of evidence used to score each tool (additional data for all 76 tools along with further information on number and categories of NTS assessed and breakdown of scores is included in Appendix 11). Where tools achieved the same score, they were ranked in publication order and if this was the same year, they were ranked alphabetically by first author. Radar plots displaying the multivariate data for the five highest and lowest scoring assessment tools are included in Figure 5-2.

Tool (Year)	Environment		Specialty / clinical arena	MDT involved plus Psych/HF	Validity		Reliability		Usability		Total score (max 27)
	Sim	Real			Response process	>2 items	internal consistency	Interrater reliability	Usability evaluated	Training specified	
TEAM[154,316,317] **(2010)**	●	●	Emergency Department MDT	●	●	●	●	●	●	●	25
OTAS[248,313,318–320] **(2004)**	●	●	Theatre MDT	●		●	●	●	●	●	24
Oxford NOTECHS[110,249,321] **(2008)**	●	●	Theatre MDT	●	●	●	●	●		●	24
ANTS[152,322,323] **(2003)**	●	●	Anaesthetists	●	●		●	●	●	●	23
Ottawa GRS[324,325] **(2006)**	●		Medical trainees (any specialty)	●	●	●	●	●	●	●	23
NOTSS[153,250,307,326] **(2006)**	●	●	Surgeons	●	●		●	●	●	●	22
ANTS-AP[267] **(2015)**	●	●	Anaesthetic practitioners	●			●	●	●	●	22
PETRA[308,327] **(2017)**	●	●	Obstetric MDT		●	●	●	●	●	●	22
UTBMNR[244,328,329] **(2004)**	●	●	Neonatal MDT	●	●		●	●		●	21
AOTP/GAOTP[256,305,330] **(2009)**	●		Obstetric MDT	●		●	●	●	●	●	21
PCST[331,332] **(2010)**		●	Paediatric cardiac surgery MDT	●	●		●	●	●	●	21
SPLINTS[309,333] **(2011)**	●	●	Scrub practitioners		●		●	●	●	●	21
TFAT[334,335] **(2011)**	●	●	Ward MDT	●		●		●	●	●	21
OSCAR[336] **(2011)**	●	●	Resuscitation MDT	●	●		●	●		●	21
WHOBARS[337] **(2016)**	●	●	Theatre MDT	●			●	●	●	●	21
CTS[315] **(2008)**	●	●	Healthcare MDT	●	●			●	●	●	20
SAFE-TeamS[338] **(2013)**	●		Students (medical, nursing)	●	●		●	●	●	●	20
ANTSdk[312,339] **(2015)**	●		Anaesthetists	●	●		●	●	●	●	20
TDRF[28] **(2002)**		●	Emergency Department MDT	●		●	●	●		●	19
Revised NOTECHS[268,340–342] **(2005)**	●		Theatre MDT	●		●	●	●		●	19
MHPTS*[252] **(2007)**	●		Healthcare MDT	●			●	●	●	●	19

T&PCM[343,344] (2008)	●		Emergency Department MDT	●		●		●		●	19
Anaesthetic trainee NTS[304,345] (2010)	●	●	Anaesthetic trainees			●	●	●		●	19
Trauma NOTECHS[346] (2012)	●	●	Trauma MDT			●	●	●		●	19
NANTSdk[306,347] (2014)	●	●	Nurse anaesthetists	●			●	●	●	●	19
T-MEX[310] (2014)		●	Students (medical) or trainee doctors				●	●	●	●	19
SWAT[348] (2015)	●	●	Surgeons on ward rounds	●		●		●	●		19
OSANTS[255] (2015)	●	●	Surgical trainees		●		●	●			19
Endo-OTAS[349] (2016)	●	●	Endovascular MDT	●					●	●	19
AeroNOTS[350] (2016)	●		Doctors in aeromedical transport		●	●		●		●	19
iTOFT[311] (2016)	●	●	Students (nursing, AHP)				●		●	●	19
Emergency Dr NTS[270,351] (2011)		●	Emergency Medicine doctors	●				●		●	18
TBR[253,352,353] (2011)	●		Intensive care MDT	●		●	●			●	18
NOTSSdk[273,354,355] (2012)	●	●	Surgeons	●			●	●		●	18
FoNTS§[356] (2013)	●	●	Foundation doctors	●			●			●	18
T-SAW-C[357] (2014)	●	●	Surgical trainees on ward rounds	●			●	●		●	18
BMS-NNTS[276] (2014)		●	Neurosurgeons	●	●			●		●	18
Healthcare Student NTS[358] (2015)	●		Students (AHP)	●			●	●	●		18
ICARS[257] (2017)	●		Surgeons in robotic surgery			●	●	●	●		18
TPOT†[359,360] (2008)	●		Healthcare MDT	●			●	●		●	17
CARDIOTEAM[361] (2010)	●		Resuscitation MDT		●			●	●	●	17
Surgical NTS[362] (2012)	●		Surgical trainees			●	●	●			17
APRC[280,363,364] (2012)		●	Trauma MDT					●	●	●	17

IPETT[365] (2013)	●		Paediatric and anaesthetic trainees	●		●	●			●	17
KidSIM[366] (2013)	●		Students (medical, nursing, AHP)				●	●		●	17
OTAS-S[271] (2014)		●	Theatre MDT	●				●		●	17
ENNTS[367] (2016)	●		Emergency Department nurses		●	●	●			●	17
MINTS-DR[314] (2017)	●		Obstetric MDT	●					●	●	17
NANTS-no[368] (2017)	●	●	Nurse anaesthetists				●	●		●	17
HPAT[369] (2002)	●		Trauma MDT		●	●			●		16

Table 5-2: Overview of assessment of environment and context of use, validity (evaluation of response process plus 2 or more additional items as described above), reliability and usability for the top 50 tools for assessment of NTS (all papers associated with tool analysis are referenced with the year of the first publication in parenthesis). The total score awarded to each tool is included.
AHP – Allied Health Professions
MDT – multidisciplinary team (this was marked on the table if more than one professional group plus additional expertise from psychologists (Psych) or human factors (HF) were involved in the design of the tool)
***The MHPTS was originally designed as a self-assessment tool but it has subsequently been used by others to design NTS tools to be used by observers and so has been included in the review.**
§FoNTS has not been peer reviewed for publication but is published as a report on NHS Scotland's website and was developed by team members involved in ANTS and NOTSS.
†TPOT is the NTS scoring system used in the TeamSTEPPS programme of team training developed by the United States Department of Defence in 2006. We have been unable to find the original data on how this system was designed and this paper was the only available source of validity and reliability data.

Additional information on the number and category of NTS assessed and the country of origin can be found in Appendix 11.

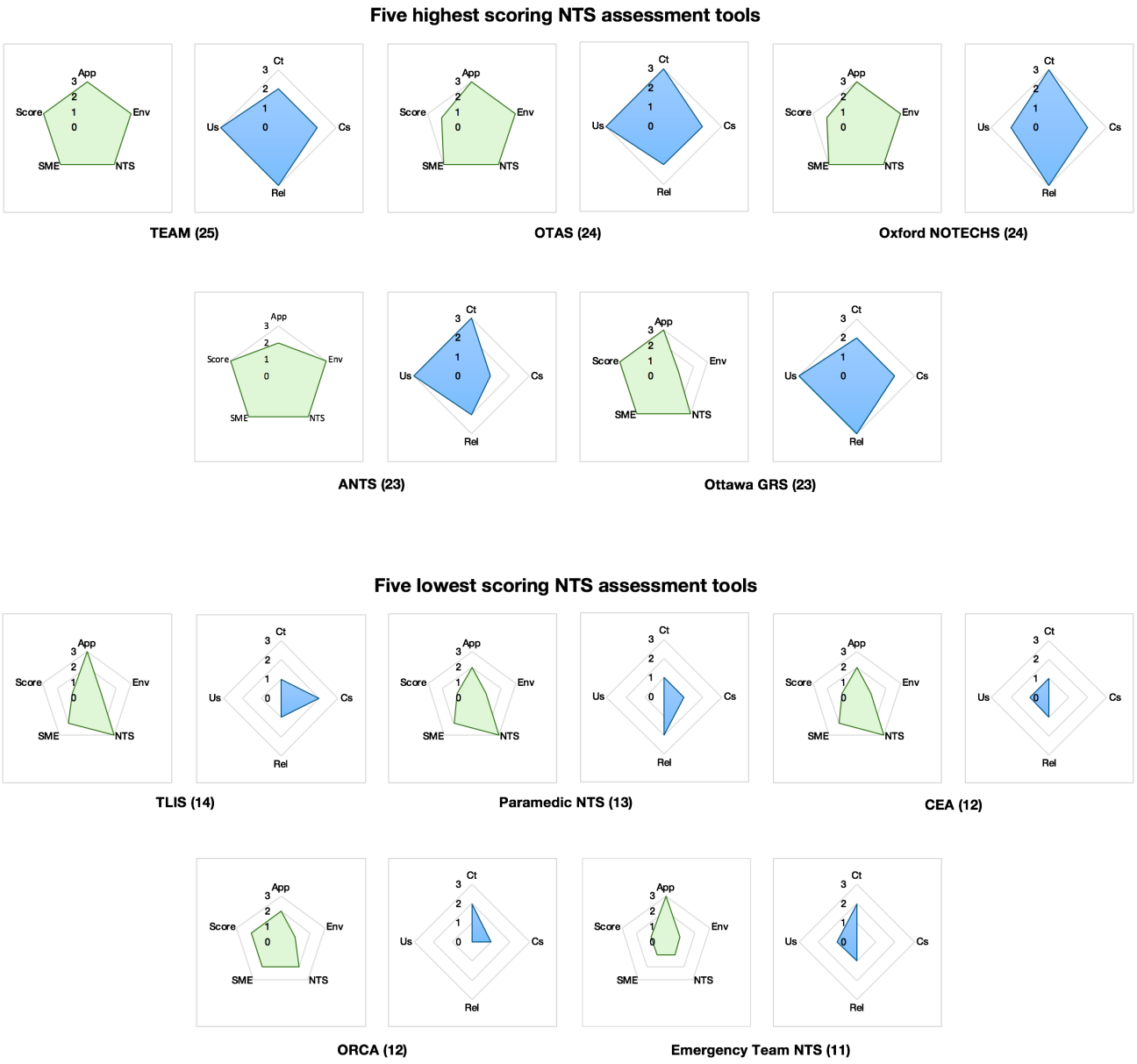


Figure 5-2: Radar plots for the five highest and lowest scoring NTS assessment tools.
Key to abbreviations: TEAM – Team Emergency Assessment Measure; OTAS – Observational Teamwork Assessment for Surgery; Oxford NOTECHS – Oxford NOn-TECHnical Skills; ANTS – Anaesthetists' Non-Technical Skills; Ottawa GRA – Ottawa Global Rating Scale; TLIS – Teamwork Leadership Interpersonal Skills; Paramedic NTS – Paramedics' Non-Technical Skills; ORCA – Operating Room Communication Assessment; CEA – Checklist of Expected Actions – Obstetric crises; Emergency Team NTS – Emergency Team Non-Technical Skills

Tool Development Score (green charts)

App – Applicability
Env – Environment
NTS – Range of NTS
SME – Subject Matter Experts
Score – Scoring system (response process)

Psychometric Testing and Usability Score (blue charts)

Ct – Content validity
Cs – Construct validity (additional items)
Rel- Reliability of tool
Us – Usability of tool

5.6 DISCUSSION

This review has provided an analysis of the growing array of NTS assessment tools in healthcare since the first was developed in 1998 by Gaba et al.[145]

5.6.1 METHOD OF DEVELOPMENT

The importance of measures which assess whole team performance has been highlighted by several authors;[31,266,370] whilst the training and assessment of individual NTS is important[46] the scoring system deliberately favoured tools which allowed more flexibility (i.e. tools for more than one profession or environment).

Instruments varied in their intended purpose, some assessed routine teamwork while others focused on management of crisis scenarios. Simulated settings allow control of scenarios and reliable depiction of behaviours (often by actors). However, it has been suggested that it is not truly representative of a real clinical environment where there may be long periods of relative calm with short bursts of intense activity, whereas a video of a simulated crisis will only focus on the 15 minutes or so of high pressure.[371] It would, therefore, seem desirable to develop tools that might be used in both settings to provide meaningful assessments during training and real clinical practice and in routine as well as emergency situations.

The NTS domains assessed were broadly similar across all the tools suggesting that they are relevant in a wide variety of clinical contexts with the appropriate context specific adaptations, which begs the question: why are there so many? Many authors stated that the reason for the development of a new tool was the lack of one relevant to their specific need. The answer may also be found, to a degree, in the necessity for compromise highlighted by Van der Vleuten[263] who described five key components in considering the utility of assessment methods:

educational impact; validity; reliability; cost and acceptability (both to examiners and examinees). He stressed that "choosing an assessment method inevitably entails compromise and the type of compromise varies for each specific assessment context" and "perfect utility is a utopia".

5.6.2 USABILITY AND TRAINING REQUIREMENTS

The issue of usability and cost of NTS assessment tools is not trivial, and has been brought into sharp relief by the current staff shortages in healthcare and difficulties in releasing staff to train. [372]

A formative training event may benefit from the use of a tool which requires little training to implement and brings additional richness to the debriefing. However, in high stakes settings evidence of validity and reliability for an assessment tool must be robust and those using it must be trained and experienced in so doing.

Most of the in-depth analysis of usability has occurred in tools developed in the past five years, suggesting a heightened awareness of the need to consider the practical use of such assessments.

The challenges of assessing NTS accurately and reliably have been enumerated by Flin[46] and Smith-Jentsch[373] and include: difficulty seeing and hearing all the relevant information; cognitive skills can be difficult to interpret and rare but important behaviours may be missed because they are not categorised. Many of the research teams who have designed these tools pointed out the challenges of using them and suggestions for best practice have been put forward by an expert group from aviation and healthcare.[69] Furthermore, Gaba,[145] Moorthy,[340] and Schraagen[331] highlight the value of simplifying the number of NTS domains analysed by a

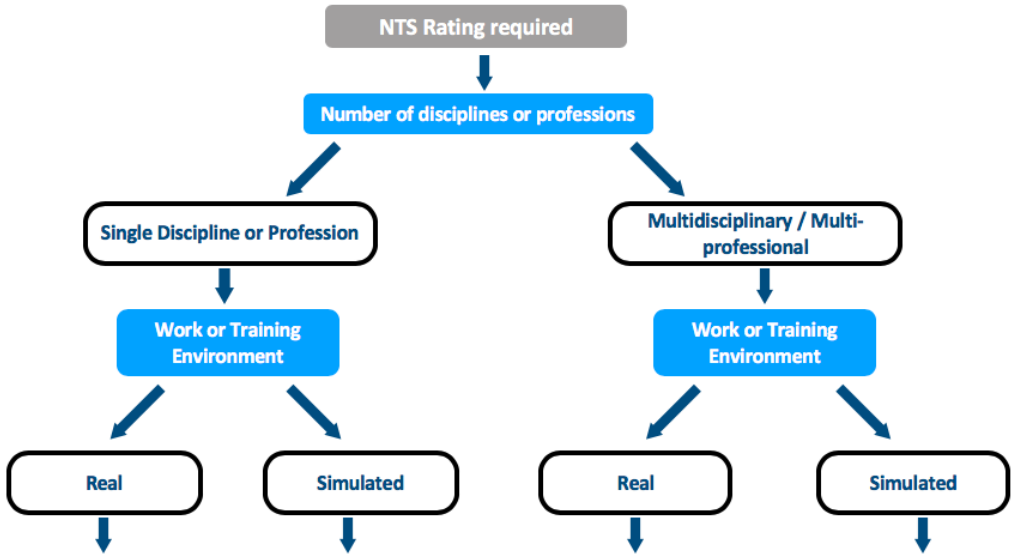
tool in order to improve the reliability of the observers. This approach may be more cost effective, Sevdalis et al showed the value of psychologist or human factors expert raters in using OTAS[320] but also recognised the resource implications. A later paper using OTAS showed that it was possible to train clinical staff to assess behaviours reliably in a short space of time.[313] Guidelines for the training of faculty in NTS assessment have since been published[70] and they stress the importance of training to ensure reliability, particularly for high stakes settings. The authors suggest a minimum requirement of two days training and a robust process of revalidation which has clear cost implications in practice.

5.7 CHOOSING AN NTS ASSESSMENT TOOL

This review has revealed the multiplicity of NTS assessment tools available in healthcare highlighting clear challenges for the educator in healthcare in trying to choose which is most appropriate for their training purposes. The process of categorising the tools in this review highlighted three initial decisions to be made:

- Is the training for a multidisciplinary team or for a single group e.g. medical students?
- Is the training in a real or simulated environment?
- What is the setting for the training e.g. ward-based, critical care or obstetrics?

Figure 5-2 provides a decision tree for considering applicability of an NTS assessment tool and all assessment tools are categorised below the tree.



Adult inpatient	SWAT [19] T-SAW-C [18] FONTS [18]	SWAT [19} T-SAW-C [18] FONTS [18] PS-OSCE-NTS [16]	TFAT [21] T-MEX [19] OTAS-WR [16]	TFAT [21]
Adult critical / emergency care	Emergency Dr NTS [18]	ENNTS [17] GRS-A [16] Trauma NTS [15] BRS-ECMO [14]	TEAM [25] OSCAR [21] TDRF [19] Trauma NOTECHS [19] EPOC [15] TTCA-24 [15]	TEAM [25] Ottawa GRS [23] OSCAR [21] Revised NOTECHS [19] Trauma NOTECHS [19] T&PCM [19] TBR [18] CARDIOTEAM [17] HPAT [16] Ottawa GRS-I [16] NTS Sepsis [15] F-TEAM [15] TLiS [14] Emergency Team NTS [11]
Obstetrics	-	CEA [12]	PETRA [22]	PETRA [22] AOTP [21] MINTS-DR [17]
Operating theatres	ANTS [23] NOTSS [22] ANTS-AP [22] SPLINTS [21] ANTSdk [20] Anaesthetic trainee NTS [19] NANTSdk [19] OSANTS [19] NOTSSdk [18] BMS-NNTS [18] NANTS-no [17]	ANTS [23] NOTSS [22] ANTS-AP [22] SPLINTS [21] ANTSdk [20] Anaesthetic trainee NTS [19] NANTSdk [19] OSANTS [19] NOTSSdk [18] ICARS [18] Surgical NTS [17] NANTS-no [17] BAR-ACRM [15] GRS-A [14] ORCA [12]	OTAS [24] Oxford NOTECHS [24] PCST [21] WHOBARS [21] Endo-OTAS [19] OTAS-S [17] OTAS-D [16] SSC and TCT [16]	OTAS [24] Oxford NOTECHS [24] WHOBARS [21] Revised NOTECHS [19] Endo-OTAS [19] OTAS-D [16]
Paediatric critical / emergency care	-	PR-NTS [16] STAT [14]	UTBMNR [21] APRC [17]	UTBMNR [21] IPETT [17]
Pre-hospital care	-	AeroNOTS [19] Paramedic NTS [13]	-	IMCBRS [16]
Undergraduate	-	L&TBM [15] OTRS [14]	iTOFT [19]	SAFE-TeamS [20] iTOFT [19] Healthcare student NTS [18] KidSIM [17] CBOATS [15] ORTAS [15]
Generic	-	-	CTS [20] CATS [16]	CTS [20] MHPTS [19] TPOT [17] CATS [16]

Figure 5-3: Decision tree for choosing a tool for the assessment of NTS in various healthcare settings. Figures in parenthesis are total scores for each tool.

5.8 STUDY LIMITATIONS

Absolute exclusion of bias is not possible but using the techniques described above can mitigate its influence.

The scoring system has flaws: tools which were published in the early days of NTS research in healthcare scored less highly due to lack of available evidence (although they were ordered according to date of publication); tools only recently published may not have had time to undertake rigorous reliability testing and tools based on those developed earlier (e.g. for use in a different language/culture) lost some marks if they relied on data from the original work. The scoring was also weighted to favour method of design over psychometric testing and usability (total score 15 versus 12 respectively). This study was not designed to assess the quality or results of psychometric or usability testing, simply if it had been undertaken. This is an area which is deserving of further analysis.

Although we contacted authors via email to ask for further information it is possible that we do not have a complete data set for each tool.

The study was restricted to considering only papers that were contiguous with the original development of the tool and did not include data from groups who had used the tools in different settings.

5.9 CONCLUSION

This review has shown that there is variability in the design and testing of the NTS tools and that consideration of these components is not always complete. Recommendations for designing and training to use tools for the assessment of NTS made by Klampfer et al[69] and

Hull et al[70] may be regarded as the gold standard but acceptability and cost implications remain a considerable barrier. Similarities between systems have also been highlighted[274,374] – strengthening support for a more unified approach to NTS teaching and a rationalisation of assessment tools.

Finally, previous reviews of NTS tools have provided an overview of available assessment techniques in different areas but have not provided a means of discriminating between them.[326,375–378]

A technique has been devised for categorising tools for the assessment of NTS and a decision tree which could be useful to both novice and expert educators in simulation based education.

The next chapter of this thesis will move on to consider the use of four observer based tools for the assessment of NTS (including SA) taken from this review. The study has been designed for the assessment of multidisciplinary teams in simulated cardiac arrest situations (in ED/ICU or theatre settings) and the tools were chosen using the decision tree described above.

CHAPTER 6 A STUDY OF RELIABILITY AND USABILITY OF NON-TECHNICAL SKILLS ASSESSMENT TOOLS FOR ANALYSIS OF VIDEO RECORDINGS OF SIMULATED CARDIAC ARREST SCENARIOS

6.1 BACKGROUND

Safe care of acutely unwell patients in dynamic clinical settings, such as the operating theatre, emergency department or intensive care unit, requires high levels of competency in both technical and NTS.

The use of experiential learning incorporating high fidelity simulation can improve competence in NTS (including SA and team-working)[31,266]. Simulation training in healthcare has expanded over the past two decades such that healthcare professionals are now trained, revalidated and assessed in simulated scenarios. Judgements made by assessors using tools designed to analyse technical and NTS must be accurate, repeatable and reliable. Chapter 5 of this thesis revealed the wide variation in clinical settings, applicability, quality of design and extent of psychometric testing for NTS assessment tools in healthcare and produced a system for scoring and choosing them.

Making an assessment of NTS at an individual level or for a whole team requires a clear understanding of the characteristics underpinning good teamwork, a knowledge of the overt behaviours which are exemplars for a particular NTS (e.g. clear communication of mental models for good SA) and what constitutes a poor, average or good performance. These are not intuitive skills and training is required to use NTS instruments reliably and to ensure that

trainers are measuring what they think is being measured. The Civil Aviation Authority describes with great clarity what is expected of its examiners and also mandates regular training and revalidation for assessors in the use of behavioural rating systems[379]. A recent expert panel put forward similar recommendations for training healthcare professionals in the use of NTS tools[380] but as yet there is no such requirement in healthcare. In this chapter four of these NTS tools have been used by three experts in simulation based education (SBE) for scoring standardised videos of cardiac arrest situations.

Overall the aim of this study was to understand if four tools which were designed in the NHS and scored well, in terms of method of development and psychometric testing in Chapter 5, could be used easily and reliably by expert raters. The more specific objectives were to:

- Assess internal consistency and inter-rater reliability of four tools for the assessment of NTS: Anaesthetic Non-Technical Skills (ANTS)[152], Observational Teamwork Assessment for Surgery (OTAS)[248], Observational Skill Based Assessment tool for Resuscitation (OSCAR)[336] and the revised Oxford Non-Technical Skills system (Oxford NOTECHS II)[381]. The choice of tools is explained below.
- Analyse the usability of ANTS, OSCAR, OTAS and Oxford NOTECHS
- Compare the reliability of assessment of SA for each tool

6.2 METHODS

6.2.1 STUDY DESIGN

Mixed quantitative and qualitative methods were used to undertake a secondary analysis of videos recorded during the START study using four NTS assessment tools. START was a study to investigate the value of simulation training for novice anaesthetists - ethics ref:

MSD/IDREC/C1/2011/137 and involved all three investigators in this study). The protocol describing a review of a subgroup of the videos using additional NTS scoring systems was submitted to the University of Oxford's Clinical Trials and Research Governance team and was accepted as a secondary review within the original terms of consent requiring no further ethical approval. Ten videos (20%) were selected randomly from a pool of 50 adult life support (ALS) scenarios. It was necessary to be pragmatic when considering the time this would take and it was agreed that 20% was an appropriate and realistic amount of the dataset to review. The videos were filmed in the University of Oxford's simulation centre (Oxford Simulation Teaching and Research – OxSTaR) and were recorded using in-situ audio-visual equipment with the consent of the study participants and stored in a secure database. The scenario was standardised and depicted an adult patient suffering with acute severe asthma who develops a tension pneumothorax and deteriorates to the point of cardiac arrest (pulseless electrical activity - PEA). The scenario was set in ED (not theatre) however, it was easy to consider the setting as an anaesthetic room or ED setting (which made it suitable for the tools we chose) and all trainees in anaesthetics work in ED environments. The candidate was expected to diagnose severe asthma and institute appropriate treatment in the first phase of the scenario and then diagnose and treat the tension pneumothorax leading to PEA according to guidelines from the UK Resuscitation Council[382]. The faculty guidelines, scenario description and algorithm for PEA arrest are included as Appendices 12 and 13.

6.2.2 PARTICIPANTS AND PROCEDURES

Three members of OxSTaR faculty with extensive experience in the assessment of NTS: Dr Nick Crabtree (NC), Dr Paul Greig (PG) and the author (HH) took part in the study. All are consultant anaesthetists with greater than ten years' experience in simulation based education and trained

in the use of the ANTS tool. The ten ALS scenarios were reviewed and the participants' NTS rated using ANTS, OTAS, OSCAR, and Oxford NOTECHS. Random numbers were assigned to the videos so that they were viewed in a different order each time for each tool. All video analyses were undertaken in environments optimised for uninterrupted viewing after familiarisation with the tools had taken place. Score sheets for each tool were marked and annotated by hand (see Appendices 14-17) and data were transcribed into a spreadsheet for subsequent analysis. Data were anonymised and stored on a university computer in a locked office.

6.2.3 NTS ASSESSMENT TOOL SELECTION

Several authors have highlighted the importance of the culture in which a tool for the measurement of NTS is to be used.[273–275,347,383,384] Therefore, four tools which had originally been developed and validated in the UK for staff in the NHS were chosen. All four tools had scored highly in the systematic review of NTS tools (please see chapter 5) and included SA as one of the domains for assessment. ANTS, OTAS, OSCAR and Oxford NOTECHS, however, displayed considerable variability in original study design, context of use and data analysis and a summary of these differences is provided in Appendix 18.

The attributes of each tool along with their scores are included in Table 6-1. The authors of OTAS, OSCAR and Oxford NOTECHS were contacted as our study would be assessing only one of the teams that they cover in their systems. All agreed that this was acceptable as the tools have been designed to provide assessment of surgical, anaesthetic and nursing teams where the construct of those teams and the context of use may differ widely.

6.2.4 MEASURES

This section provides an overview of the characteristics of the four assessment tools; how the investigators trained to use the assessment tools; how the method of scoring was standardised and the how usability of the tools was analysed.

6.2.4.1 NTS ASSESSMENT TOOL CHARACTERISTICS

After initial agreement that each of the tools was feasible to be used for a post hoc review of video data of simulated resuscitations, tool characteristics were considered under the headings of tool development, psychometric testing and usability and data are summarised from Chapter 5 in Table 6-1.

NTS Tool (Date of publication)	Tool development (see chapter 5)	Psychometric testing and usability (see chapter 5)
ANTS: Anaesthetic Non-Technical Skills (UK 2003)	Applicability: anaesthetists only Environment: real and simulation Range of NTS: 4 categories, 15 elements SME: MDT, included psychology Scoring system: described, justified and compared with others **Score: 14/15**	Validity: content and construct Reliability: IRR (r_{WG}: 0.56-0.65) and internal consistency (Cronbach's α for elements in each category 0.79-0.86) Usability: quantitative and qualitative assessment; bespoke training programme described, handbook available **Score: 9/12**
OTAS: Observational Teamwork Assessment for Surgery (2004)	Applicability: ratings for surgical, anaesthetic and nursing staff in operating theatre during pre-, per- and post-operative phases, aggregate score for whole team Environment: real and simulation Range of NTS: 5 categories, 11 elements SME: MDT, included psychology Scoring system: described and justified **Score: 14/15**	Validity: content and construct Reliability: IRR (ICC: $\geq$ 0.68) Usability: quantitative and qualitative assessment; bespoke training programme described, handbook available **Score: 10/12**
OSCAR: Observational Skill Based Assessment tool for Resuscitation (2011)	Applicability: ratings for anaesthetic, physician and nursing staff in resuscitation settings Environment: real and simulation Range of NTS: 6 categories, 48 elements SME: MDT, included psychology Scoring system: described, justified and compared with others **Score: 15/15**	Validity: content and construct Reliability: IRR (ICC: 0.65-0.91); internal consistency (Cronbach's α: 0.74-0.97) Usability: no assessment of usability; training through self-study and group practice sessions **Score: 6/12**
Oxford NOTECHS II: Revised Oxford Non-Technical Skills Score (2008)	Applicability: ratings for surgical, anaesthetic and nursing teams in operating theatre Environment: real and simulation Range of NTS: 4 categories, 16 elements SME: MDT, included psychology and human factors experts Scoring system: described and justified **Score: 14/15**	Validity: content and construct Reliability: IRR (ICC 0.34-0.88); test-retest Usability: qualitative assessment; training through self-study and group practice sessions **Score: 10/12**

Table 6-1: Characteristics of four Non-Technical Skills Assessment Tools for Healthcare Professionals (MDT = multidisciplinary team; IRR = interrater reliability; ICC = intraclass correlation) *Oxford NOTECHS II describes the revised version of Oxford NOTECHS which led to a refinement of the scoring system (from a 6 to an 8 point scale) but no change to the NTS domains assessed. The tool will be referred to as Oxford NOTECHS for the purposes of this study.

6.2.4.2 TRAINING IN THE USE OF THE ASSESSMENT TOOLS

Training in the use of the remaining tools was devised through discussion with the tool's original designers (Professor N. Sevdalis for OTAS and OSCAR and Professor P. McCulloch and Dr L. Morgan for Oxford NOTECHS all provided helpful advice) and providing them with an explanation of the experience of the investigators and the design of the study. To ensure consistency of training and commonality of approach NC, PG and HH read the materials provided and then reviewed five randomly assigned ALS videos together using ANTS, OTAS, OSCAR and Oxford NOTECHS with discussion of scoring differences and the nuances of use of the tools.

During the initial stages of familiarisation with the tools it became clear that OTAS would not be suitable for use in this study. The OTAS tool was designed for the assessment of multidisciplinary healthcare professionals in the operating theatre during three distinct stages: the pre-, intra- and postoperative periods. During preliminary review of the OTAS handbook it was felt that OTAS could be used as if the cardiac arrest had occurred whilst the patient was in the anaesthetic room. However, because OTAS scoring focused on key steps during anaesthesia and surgery, using OTAS solely for an arrest event would not allow the full range of behaviours to be observed and would lead to a falsely low score. We, therefore, restricted further analysis to the use of ANTS, Oxford NOTECHS and OSCAR.

6.2.4.3 SCORING SYSTEMS FOR ANTS, OXFORD NOTECHS AND OSCAR

ANTS, Oxford NOTECHS and OSCAR divide NTS into different categories and elements (I have deliberately chosen the ANTS taxonomy here for simplicity and consistency), score individuals or teams and use different scoring systems. To compare the scales, it was necessary to standardise

the way in which we assessed our data at the element and category level. A summated score of the categories in ANTS was added (as this is normal practice for Oxford NOTECHS and OSCAR) and element scores were recorded for Oxford NOTECHS as this is normal practice for ANTS and OSCAR. A comparison is provided in Table 6-2. Scores for Oxford NOTECHS were only recorded for the anaesthetic team as there was no surgical or nursing team and for OSCAR only a physician team score was recorded as there was no anaesthetic or nursing team.

System and profession(s) assessed	Categories	Number of elements	Score
ANTS **Rating for anaesthetist only**	Task management	4	1-4 (plus "not observed") for elements, overall category score No global score for all categories Summated global category scores added (20 scores per video)
	Team working	5	
	Situation awareness	3	
	Decision making	3	
	Total = 4	Total = 15	
Oxford NOTECHS **Rating for three theatre teams – surgical [S], anaesthetic [A] and nursing [N], all use the same categories and elements**	Leadership and management (S,A,N)	5	1-8 for categories Includes global summated score for categories Score for each element added (21 scores per video)
	Teamwork and co-operation (S,A,N)	4	
	Problem solving and decision making (S,A,N)	4	
	Situation awareness (S,A,N)	3	
	Total = 4	Total = 16	
OSCAR **Rating applied to anaesthetic group [A], physician group [P] and nursing group [N], elements specific to each group**	Communication	A=4,P=3,N=3	0-6 for elements and categories Includes global summated score for categories (25 scores per video)
	Co-operation	A=2,P=2,N=3	
	Co-ordination	A=2,P=2,N=3	
	Leadership	A=3,P=3,N=2	
	Monitoring (SA)	A=3.P=3,N=2	
	Decision making	A=3,P=2,N=3	
	Total = 6	Total = 32	

Table 6-2: Differences in structure (including number and type of category and number of elements) and scoring of ANTS, Oxford NOTECHS and OSCAR; adaptations for this study are highlighted in red.

6.2.4.4 USABILITY OF ANTS, OXFORD NOTECHS AND OSCAR

Quantitative and qualitative assessments of the usability of ANTS, Oxford NOTECHS and OSCAR were made.

Quantitative measures:

- Time taken to train to use the assessment tools (including reading and assimilating Information; meeting to assure consensus on use of the tools; and a group training and familiarisation session)
- Completeness of data points filled for each system

  Time taken to review and score the videos using each assessment tool (measured for HH as NC and PG had been involved in original review of the START videos)
- Quantitative data from the usability questionnaire (see below)

Qualitative measures:

- Questionnaire adapted from the usability assessment devised for the development of the ANTS-AP NTS[267] assessment system (kindly provided by Dr J. Rutherford – please see Appendix 19). Questionnaires were answered independently by each investigator at the training session and then again after assessment of the ten study videos
- Post-study meeting to discuss tool attributes. Results from written questionnaires and discussions were combined

6.3 DATA ANALYSIS

Scores for each system (global and category) were assessed for normality of distribution and are displayed as raw and percentage scores. Comparisons of descriptive statistics were made between global scores for each system (i.e. all categories combined) and between the SA categories, as all the assessment tools used three elements to score SA (none of the other categories had the same number of elements contributing to the score).

Reliability of the assessment tools was analysed using Cronbach's alpha for internal consistency across all raters for global scores and category scores in each tool. The statistical tests used to calculate interrater reliability (IRR) were those described in the original papers for these tools and in others described in Chapter 5: weighted (Cohen's) kappa, Intraclass Correlation Coefficients, and within groups reliability scores (r_{WG}).

The weighted kappa can only be used to compare two raters, therefore, we randomly allocated one pair for this analysis (PG and HH) and calculated ICC and r_{WG} for all three raters and PG and HH alone.

When the within groups reliability score (r_{WG}) was applied to our data it revealed very high levels of agreement in all categories for all raters (i.e. it did not discriminate at all between NC, PG and HH). Advice from an expert statistician was to exclude r_{WG} from the analysis as it is subject to significant test bias and benchmarking is only possible for tests using a score range equal to or greater than five with 10 raters or more. Interrater reliability was, therefore, calculated with ICC and weighted kappa only for the overall NTS tool scores and for each of the NTS categories.

Time taken to assess videos using each tool was compared using one-way ANOVA (SPSS V24.0).

6.4 RESULTS

6.4.1 SCORES USING ANTS, OXFORD NOTECHS AND OSCAR

Data were analysed in SPSS (V24.0) and revealed that global scores and category scores were not normally distributed (Shapiro-Wilk <0.05) therefore, Table 6-3 shows the median, range and

interquartile range (IQR) for global and category scores for each rater using each system. Table 6-4 shows the scores as percentages of the possible total for each tool.

Scoring system and NTS category	HH Median (range) [IQR]	NC Median (range) [IQR]	PG Median (range) [IQR]
ANTS (1-4)			
Global score	12.5 (8-16)[4]	16.0 (14-16)[1]	12.0 (9-14)[3]
Task management	3.0 (2-4)[0]	4.0 (4, 4)[0]	3.0 (2-3)[1]
Teamwork	3.5 (2-4)[1]	4.0 (3-4)[0]	3.5 (2-4)[2]
Situation Awareness	3.0 (2-4)[0]	4.0 (3-4)[0]	3.0 (1-3)[0]
Decision making	3.0 (2-4)[1]	4.0 (3-4)[0]	3.0 (3-4)[0]
Oxford NOTECHS (1-8)			
Global score	27.0 (16-31)[7]	28.5 (22-32)[7]	25.0 (16-29)[6]
Leadership and management	7.0 (5-8)[2]	7.0 (5-8)[2]	6.0 (4-7)[2]
Teamwork and cooperation	6.5 (4-8)[2]	7.0 (6-8)[2]	6.0 (3-8)[2]
Problem solving and decision making	6.5 (3-7)[2]	7.0 (6-8)[2]	6.0 (4-7)[1]
Situation Awareness	7.0 (4-8)[1]	6.5 (5-8)[2]	6.0 (5-7)[1]
OSCAR (0-6)			
Global score	31.0 (20-34)[10]	26.5 (23-36)[11]	25.5 (13-34)[9]
Communication	5.0 (4-6)[1]	4.5 (3-6)[1]	4.0 (2-6)[2]
Cooperation	5.0 (3-6)[2]	4.0 (3-6)[1]	4.0 (2-5)[2]
Coordination	5.0 (4-6)[1]	4.0 (4-6)[2]	4.0 (2-5)[2]
Leadership	5.0 (3-6)[1]	4.5 (4-6)[2]	4.0 (2-6)[1]
Situation Awareness	5.0 (3-6)[2]	4.5 (4-6)[2]	4.5 (3-6)[1]
Decision making	5.0 (3-6)[2]	5.0 (4-6)[2]	4.0 (3-6)[1]

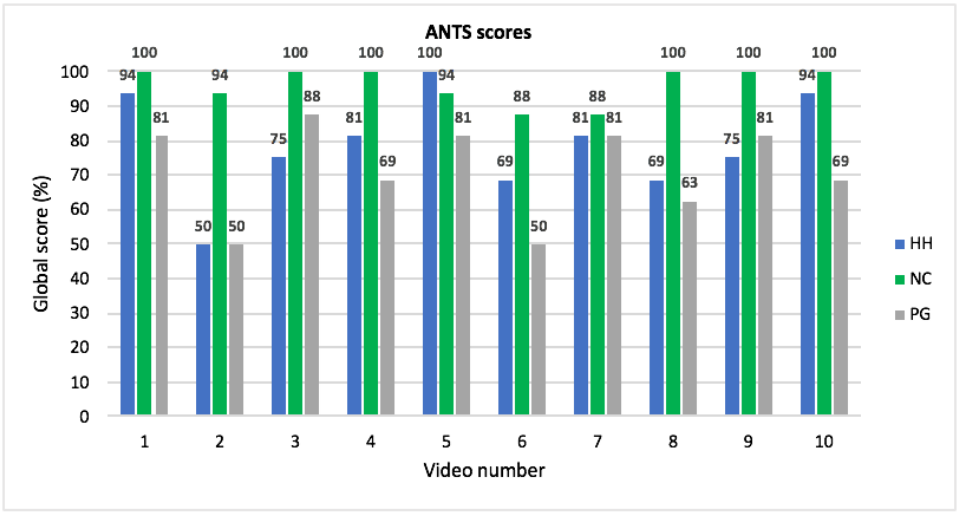
Table 6-3: Median (with range and IQR) raw scores (overall and for each category) for ANTS, Oxford NOTECHS and OSCAR; score ranges for each system are shown in parenthesis

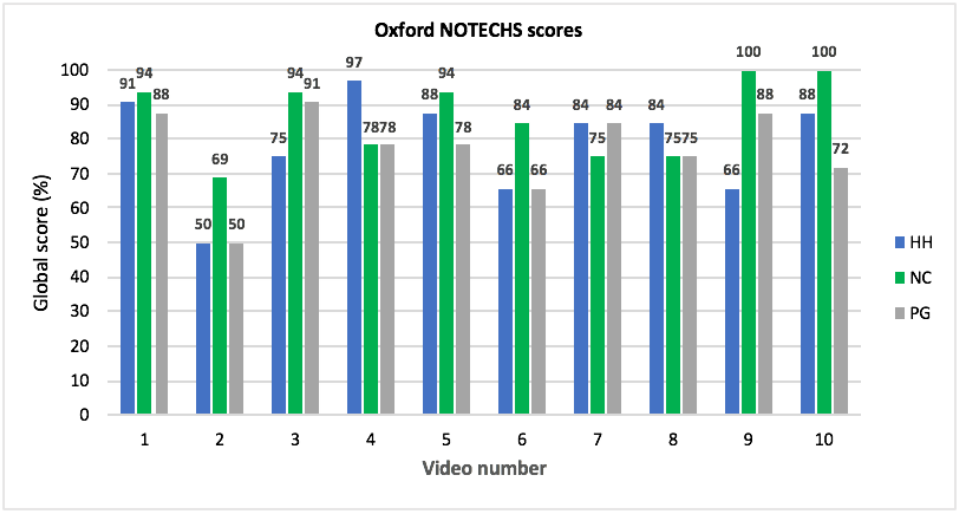
Median scores for performances were above average for all raters and all systems suggesting that the teams in the videos were generally performing well. Median ANTS scores for NC were a maximum four points suggesting a ceiling effect was evident for this rater with this tool. Percentage scores were calculated to allow comparison across the different assessment tools (see Table 6-4 and Figure 6-1). The "conflict solving" element of the teamwork and cooperation category for Oxford NOTECHS was not relevant in the context of the ALS scenario and so was removed from the analysis.

Scoring system and NTS category	HH %	NC %	PG %
ANTS			
Global score	78	100	75
Task management	75	100	75
Teamwork	87.5	100	87.5
Situation Awareness	75	100	75
Decision making	75	100	75
Oxford NOTECHS			
Global score	84	89	78
Leadership and management	87.5	87.5	75
Teamwork and cooperation	81	87.5	75
Problem solving and decision making	81	87.5	75
Situation Awareness	87.5	81	75
OSCAR			
Global score	74	63	61
Communication	71	64	57
Cooperation	71	57	57
Coordination	71	57	57
Leadership	71	64	57
Situation Awareness	71	64	64
Decision making	71	71	57

Table 6-4: Percentage global and category scores for ANTS, Oxford NOTECHS and OSCAR for each of the three raters.

Percentage scores revealed differences between raters and assessment tools but global scores for OSCAR were the lowest for all raters. Table 6-4 provides a breakdown of scores for each video and each assessment tool. The lowest scoring video when scores were averaged across raters was video two. The interesting discrepancy is the ANTS score given by NC which remained high.





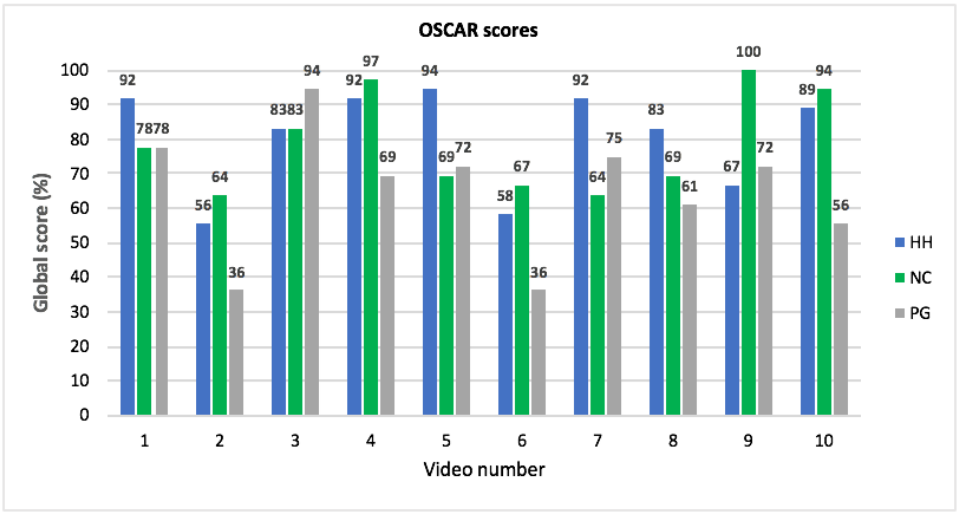


Figure 6-1: Percentage scores for each video scored independently by three raters using ANTS, Oxford NOTECHS and OSCAR

6.4.2 RELIABILITY OF ANTS, OXFORD NOTECHS AND OSCAR

6.4.2.1 INTERNAL CONSISTENCY

Cronbach's alpha was used to calculate internal consistency for all raters using all assessment tools (see Table 6-5). Results were good (a score of ≥0.7 is considered satisfactory[385]) across all categories combined and for individual categories when all raters were combined. Scores are highlighted in red where they fall below 0.7 and this happened for NC and PG mainly for ANTS, the most familiar system.

Scoring system and NTS category	HH	NC	PG	Combined
ANTS				
All categories	0.95	0.80	0.92	0.95
Task management	0.71	0.20	0.51	0.76
Teamwork	0.90	0.87	0.93	0.92
Situation Awareness	0.90	0.60	0.70	0.77
Decision making	0.70	0.20	0.54	0.74
Oxford NOTECHS				
All categories	0.99	0.96	0.97	0.98
Leadership and management	0.96	0.89	0.93	0.93
Teamwork and cooperation	0.95	0.89	0.95	0.94
Problem solving ,decision making	0.96	0.78	0.94	0.91
Situation Awareness	0.97	0.94	0.72	0.93
OSCAR				
All categories	0.97	0.97	0.98	0.97
Communication	0.86	0.88	0.82	0.83
Cooperation	0.77	0.77	0.68	0.76
Coordination	0.80	0.89	0.96	0.92
Leadership	0.85	0.93	0.89	0.93
Situation Awareness	0.91	0.84	0.91	0.88
Decision making	0.91	0.86	0.96	0.90

Table 6-5: Cronbach's alpha for scores from each rater for ANTS, Oxford NOTECHS and OSCAR; Scores in red fall below the acceptable level of reliability for summative settings

6.4.2.1.1 INTERNAL CONSISTENCY FOR SA CATEGORY

Cronbach's alpha was ≥0.7 for all raters apart from NC with ANTS where a score of 0.6 was recorded.

6.4.2.2 INTERRATER RELIABILITY

Inter-rater reliability was calculated in SPSS using ICC and weighted kappa and results comparing the three investigators using each tool are shown in Table 6-6.

Scoring system and NTS category	ICC: PG and HH (all 3 raters)	Weighted kappa (PG and HH only)
ANTS		
Global score	0.73 (0.62)	0.52
Task Management	0.26 (NA)	0.10
Teamwork	0.65 (0.62)	0.64
Situation Awareness	0.79 (0.60)	0.54
Decision Making	0.64 (0.67)	0.45
Oxford NOTECHS		
Global score	0.71 (0.69)	0.54
Leadership and Management	0.46 (0.50)	0.28
Teamwork and Cooperation	0.77 (0.76)	0.61
Problem Solving and Decision Making	0.81 (0.67)	0.67
Situation Awareness	0.34 (0.51)	0.22
OSCAR		
Global score	0.80 (0.68)	0.40
Communication	0.53 (0.25)	0.25
Cooperation	0.84 (0.69)	0.54
Coordination	0.72 (0.75)	0.26
Leadership	0.75 (0.67)	0.29
Situation Awareness (monitoring)	0.87 (0.67)	0.73
Decision making	0.64 (0.66)	0.41

Table 6-6: IRR results for all raters or paired raters (PG and HH) using ICC, and weighted kappa. Scores are highlighted in green where good or better agreement occurred and underlined where p=<0.05

The ICC results show good or better agreement using benchmarking described by Downing[141] when PG and HH are compared for all global scores and for most categories in OSCAR, the teamwork and problem solving categories in Oxford NOTECHS and the SA category for ANTS.

Altman's[386] updated version of the Landis and Koch[387] benchmarking system was used to judge results for the kappa statistic results (a score >0.6 indicates good agreement).

6.4.2.2.1 INTERRATER RELIABILITY OF THE SA CATEGORY

The IRR for SA was good for HH and PG using ANTS and OSCAR but poor for Oxford NOTECHS for PG and HH and all raters combined. This is interesting because ANTS and Oxford NOTECHS describe SA in very similar ways i.e. by referring to perception, comprehension and projection, whereas OSCAR provides specific behavioural exemplars (e.g. checks time, reassesses protocol) with no reference to the recognised levels of SA.

6.4.3 USABILITY OF ANTS, OXFORD NOTECHS AND OSCAR

6.4.3.1 QUANTITATIVE MEASURES

6.4.3.1.1 TRAINING TIME

The initial period of familiarisation with the two tools which had not previously been used by the investigators comprised three hours reading the original papers for OSCAR and Oxford NOTECHS and reviewing their scoring systems followed by a four hour session of video reviews and discussion using ANTS, Oxford NOTECHS and OSCAR as described above.

6.4.3.1.2 COMPLETENESS OF DATA COLLECTION

Data were complete on all score sheets for all systems and all raters.

6.4.3.1.3 TIME TAKEN TO SCORE VIDEOS

Times taken to complete scoring of the videos by HH using ANTS, OSCAR and Oxford NOTECHS are shown in Table 6-7: Mean times in minutes (with 95% CI) for single rater (HH) to review videos with ANTS, OSCAR and Oxford NOTECHS

Times include the length of the video. Data were tested for normality (times were normally distributed as assessed by Shapiro-Wilks score: p>0.05) and mean times in minutes with their 95% confidence intervals are shown in Table 6-7.

Scoring system	Mean times - minutes (95% CI)
ANTS	15.3 (13.8-16.7)
Oxford NOTECHS	18.5 (16.6-20.5)
OSCAR	19.6 (17.7-21.4)

Table 6-7: Mean times in minutes (with 95% CI) for single rater (HH) to review videos with ANTS, OSCAR and Oxford NOTECHS

The one-way ANOVA test was applied to compare times taken to use each of the assessment tools and revealed that time taken to use ANTS was significantly lower than Oxford NOTECHS (p=0.02) and OSCAR (p=0.002) but there was no significant difference between Oxford NOTECHS and OSCAR.

6.4.3.1.4 USABILITY QUESTIONNAIRE

OSCAR scored lowest across questions relating to behaviours described (questions 3,4,5,9,15) and ease of use (questions 6,7,10,11,12) when compared with ANTS and Oxford NOTECHS. One rater (NC) felt that more information for training to use OSCAR was necessary. The final question (16) on overall usability was also negative from all raters for OSCAR. Complete answers are provided as Appendix 20.

6.4.3.2 QUALITATIVE MEASURES

Qualitative data from the usability questionnaires and the subsequent review meeting are summarised here with quotes taken from written or verbal transcripts (all qualitative data are summarised in Appendix 21).

Comments about the systems overall highlighted the differences in context of use: ANTS "can only be used to score the anaesthetist in the team" whereas Oxford NOTECHS and OSCAR "assess three sub-teams" although it was pointed out that this would require additional context specific expertise from the faculty. Oxford NOTECHS was found to be easier to use than OSCAR because of its similarity of construct to ANTS.

The rating scales for Oxford NOTECHS and OSCAR were preferred to ANTS because it was felt less likely that a ceiling effect would be observed. However, one sided assessment sheets were preferred and OSCAR's three page layout was considered unwieldy.

6.5 DISCUSSION

This study has explored the use of three different tools to assess NTS in context of a standardised simulated emergency scenario. A similar study considering three different tools[294] (TEAM[154], T-NOTECHS[346] and TTCA[294]) in which three raters assessed 10 non-standardised videos of real trauma care episodes (five emergency and five non-emergency) using the tools. The three raters trained to use the less familiar tools (TEAM and T-NOTECHS) in a similar way to this study and found a variation in IRR (using ICC) which resonated with our findings. The results of the study also highlighted issues in scoring, internal consistency, and usability of the tools.

6.5.1 SCORES USING ANTS, OXFORD NOTECHS AND OSCAR

The three raters in this study are all accustomed to using NTS tools (usually ANTS) in formative debriefing settings and rarely ascribe numerical scores to candidates or teams. Performance is considered in the context of what has just played out in the simulator and debriefing uses verbal descriptors and objective examples (either remembered or recorded) of performance to enhance learning in a supportive environment.[388,389] Ratings of NTS where there is more than

one member of faculty are usually derived for categories and global scores by consensus, prior to the debrief beginning and interrater reliability scores are not relevant because complete agreement is reached. Whilst providing a score as a marker of performance is important in discriminating between levels of performance and as a means of calculating IRR Flin et al[390] do not recommend the use of scores for formative debriefing.

The assessment tools used in this study all provided different scoring systems. The score range was lowest for ANTS (1-4) and highest for Oxford NOTECHS (1-8) and a ceiling effect was apparent in NC's scores for ANTS. The lack of variance for NC's scores in ANTS may have affected the IRR results. We scored the same scenario (with different candidates) for each of the tools to provide some standardisation of expected actions and behaviours. It is interesting to note that both tools which used videos to test IRR (ANTS and OSCAR) did not do this.

The majority of our data were not normally distributed which is why median values have been displayed as a measure of the central tendency of the scores for each rater. All of the original papers discuss mean values with no mention of the distribution of their scores. This is important when one considers that the research group who designed Oxford NOTECHS originally described a 6 point rating scale[249] but later adapted it in recognition of the fact that this did not allow enough discrimination between candidates or teams. However, the revised score range for Oxford NOTECHS II (1-8) suggests a starting point of 6 (assuming that most teams will perform to an acceptable level) which automatically skews scores to the top end of the scale.

6.5.2 RELIABILITY OF ANTS, OXFORD NOTECHS AND OSCAR

6.5.2.1 INTERNAL CONSISTENCY

Cronbach alpha scores for all categories with all raters combined were good for all tools. However, when raters were considered separately the assessment tool with the lowest score for internal consistency was ANTS (this was even more obvious when categories were considered separately). It is possible that this is because less time was taken to consider how each of the raters used ANTS as it was the tool we had most experience with. The Civil Aviation Authority[379] require that trainers' performance is regularly reviewed and these results suggest that it is also important to re-examine raters' use of assessment tools (of any form) in healthcare settings.

6.5.2.2 INTERRATER RELIABILITY

Measurements of particular attributes in the same subjects may vary greatly between raters and this source of unpredictability is an obvious concern in clinical settings but also for examinations, particularly in high stakes settings. This fact is further complicated because many measurements between raters ignore the presence of rater variance and assume that differences are caused by a change in the attribute being assessed, whether that is a clinical sign or a behaviour.[391]

The challenge in comparing reliability of NTS assessment tools in healthcare is magnified by the variety of different scores analysed (e.g. means, raw scores or global scores) and statistical tests used by the developers. The majority of studies of NTS assessment tools in chapter 4 used ICC or kappa (usually weighted) but a few used r_{WG} or generalisability theory. The choice of statistical

assessment in this study was governed by relevant literature;[141,391,392] expert statistical advice and by tests which had been used in the original studies. Two tests (ICC and weighted kappa) were chosen to analyse the same data and provide an opportunity to highlight the ease with which reliability may be misinterpreted.

This study showed that the ICC scores of three expert raters using three different NTS assessment tools for the analysis of 10 standardised videos ranged from poor (task management in ANTS and SA in Oxford NOTECHS) to very good (problem solving in Oxford NOTECHS and cooperation and SA in OSCAR). ICC is recommended as the test to use for IRR by Gwet[391] (personal communication: "I always first recommend the use of ICC with quantitative (i.e. numeric) measurements regardless of the number of judges") and Downing[141] and ICC results were good to very good for HH and PG in 9 of the 15 categories and all the global scores. However, the weighted kappa results showed only fair agreement in 7 of the 15 categories and moderate agreement for the global scores for all tools.

The IRR for ANTS was surprisingly moderate despite the calibration session prior to rating the videos individually. As outlined above, we were not as explicit about assessing particular elements as was the case with Oxford NOTECHS and OSCAR. Furthermore, whilst each of the three raters is regularly using NTS assessment in their debriefing sessions they do not routinely do so together and do not formally score participants.

IRR was better for OSCAR than for Oxford NOTECHS (when assessed with ICC) which came as a surprise because Oxford NOTECHS is more similar in structure to ANTS. OSCAR, however, provides more explicit example behaviours (because it is only considering NTS in one clinical situation: cardiac arrest) within the categories which may have reduced variance between raters.

Some authors have recommended generalisability theory as the most comprehensive assessment of sources of variance in studies of reliability.[141,270,310] Generalisability theory requires substantial numbers of raters and subjects, however, and our study was not large enough to produce meaningful results from such an analysis.

Finally, IRR scores for the SA category were very good for ANTS and even better for OSCAR but poor for Oxford NOTECHS. Several of the studies describing NTS tools reference SA as being a challenging category to score.[152,153,276,351] It is possible that, in this study, our familiarity with ANTS and the prescriptiveness of OSCAR led to better scores as well as the bias of our interest in research into SA. The lower score for Oxford NOTECHS may relate to the difficulty we all mentioned with the scoring system in the qualitative analysis (see appendix 21).

6.5.3 USABILITY OF ANTS, OXFORD NOTECHS AND OSCAR

6.5.3.1 QUANTITATIVE ASSESSMENT OF USABILITY

Other studies have used completeness of score sheets as a marker of usability of a system[275,393,394] but it is a crude method. The completeness of the score sheets in this study masked the underlying issues with the tools which were elucidated in the qualitative data (see below).

Time taken to assess the tools was measured as described above. There was no significant difference between time taken to use Oxford NOTECHS and OSCAR despite the fact that OSCAR had more categories and elements to assess. This may be explained by the fact that OSCAR provides explicit guidance for each of its elements, even though the score sheet covers several pages. The shorter time taken to assess with ANTS is explained in part by familiarity with the tool (although the score sheet is rarely used in the author's practice) and by the fact that only

one team member is being assessed. It would be interesting to analyse the difference in time taken when all three teams are assessed using Oxford NOTECHS and OSCAR but that was not possible in this study.

6.5.3.2 QUALITATIVE ASSESSMENT OF USABILITY

Whilst statistical evidence of reliability provides useful information about the validity of a tool it does not complete the picture.[245] The analysis of usability highlighted some important differences between the tools which would impact our choice of tool in future studies and are highlighted in Appendix 21. All raters felt that observing behaviours relevant to categories and elements was average to easy with ANTS and Oxford NOTECHS but not with OSCAR and all felt that there were some behaviours missing from OSCAR and that descriptors of behaviours (either good or bad) were not helpful. This may have been because we were only using OSCAR to score the physician group but in our post-study meeting we all agreed that the problem stemmed from overlap of behaviours between the physician and anaesthetic group and disagreement about some of the descriptors based in our own clinical experience. This study did not allow an assessment of the use of the tool with all groups present and is one of the limitations highlighted above.

Rating scales and scoresheets also differed between the tools and the design, particularly of the OSCAR sheet caused some challenges in marking videos because it filled several sheets requiring the rater to flip between sections when different behaviours were observed. Both the Oxford NOTECHS and OSCAR sheets did not provide enough room for comment which would have been compounded if all teams were being observed. Rating behaviours is also challenging when one does not have the necessary context-specific expertise i.e. an anaesthetist would find it more

difficult to rate the behaviour of a surgical scrub-nurse because the schemata required to judge how a candidate was responding would not be accurate. Both Oxford NOTECHS and OSCAR are designed to be used to assess some or all of the teams included by one or more raters with "limited instruction" and the original studies show that IRR was acceptable for raters without clinical experience. Guidance on the use of behaviourally anchored rating scales highlights the need for extensive training in their use (especially for high-stakes settings), that they do not apply across domains and cultures (i.e. aviation to medicine, doctor to nurse) and that understanding of the context of application is vital.

6.5.3.3 TRAINING TO USE NTS ASSESSMENT TOOLS

The issue of training to ensure adequate IRR has been raised by several authors.[70,395–397] The designers of ANTS[152] are very clear on the need to provide adequate training for the use of the tool and have designed a two-day bespoke course complete with handbook whereas the authors of Oxford NOTECHS[249] state "The scale can also be used by an observer from a variety of backgrounds, with a small provision for training" and OSCAR[336] "the user would require some limited instruction in its use." In this study we undertook the training suggested by the authors and found that we did not achieve excellent reliability. Russ et al[313] describe using 8-10 videos to achieve satisfactory reliability for novice assessors and Spanager et al[355] found that experienced trainers could achieve good reliability with five. However in a larger study Graham et al[398] found that reliability was moderate to poor for a group of experienced anaesthetists trained to use ANTS in one day, which is more in line with our findings. Our less than perfect agreement may be explained in part by the lack of time spent recalibrating for ANTS and not enough time to get used to Oxford NOTECHS and OSCAR although all raters felt that the training

and material were adequate. Patey et al[399] point out the importance of training and also refreshing skills in NTS assessment and that substantial barriers exist for educators in healthcare in accessing the necessary training.

6.6 STUDY LIMITATIONS

The three expert raters in this study are more familiar with the ANTS assessment tool than Oxford NOTECHS or OSCAR. To mitigate for this we produced a standardised questionnaire which had been validated for use in the assessment of NTS tools[267] to provide an objective assessment of the different systems.

We assessed the three tools in this study by asking three expert raters to review 10 standardised videos with each tool i.e. a total of 30 videos each. Gwet[391] has highlighted that the accuracy with which one can interpret IRR results (for any test used) is dependent on number of subjects, number of raters and number of categories scored. The higher the number of subjects, raters and categories the more likely the output from the agreement statistic is to be accurate and, therefore, meaningful. This study would have benefitted from the use of more raters or a larger sample size but the design was pragmatic in the context of the time available.

Capturing and recording assessments of NTS in a scenario depicting changes happening over a short period of time is challenging (even using post-hoc video review) and it is possible that our raters missed or misinterpreted behaviours leading to an inaccurate score[68,400].

There were three people forming the team in each scenario – two of them were faculty members who were the nurses in the room and whilst this may have detracted from the realism of the situation, they had been primed to respond as they would in real life to the anaesthetist

leading the management of the cardiac arrest. We used a standardised cardiac arrest scenario where the expected team responses and actions were the same each time in order to reduce the impact of an additional source of variability on the results of the assessments.

It is possible that observer bias impacted our results as both NC and PG had been more closely involved in running the START study than HH. However, the study was completed over two years before this analysis and so clarity of memory of the scenarios was reduced. Furthermore, we randomised the videos we chose from the 50 available and also randomised the order of viewing when using each tool[401,402].

Social desirability bias may have been present in the candidates in the scenarios i.e. they may have been behaving in a manner they believed was expected of them in a simulation and this may not represent completely how they would act in the real world. This phenomenon is difficult to remove completely from a simulated setting but may be ameliorated by putting candidates in scenarios with people they usually work with or by using in-situ simulation neither of which was possible for this study.

Only one of the tools in the study was specifically designed for the measurement of NTS in resuscitation (OSCAR) – the other two, however, were designed for the assessment of NTS in anaesthetists in elective and emergency settings (including cardiac arrest). Furthermore, OSCAR and Oxford NOTECHS do not provide the option to record that a behaviour was not observed which could lead to a falsely low score in an otherwise highly performing team. In our study the "resolves conflict" element of the problem solving and decision making category in Oxford NOTECHS was not observed and was not included in the analysis.

6.7 CONCLUSION

The results from this study resonate with the challenges faced in analysing and comparing NTS assessment tools in the systematic review described in the preceding chapter. Three expert raters could not provide excellent IRR when using three different NTS tools, although reliability for the SA domain using ANTS was good and using OSCAR, very good. The next chapter will consider the analysis of SA in real-time simulations of crisis scenarios on AICU and will use ANTS because the IRR score and the usability score for the tool combine to make it the best option for the observer-based measure of SA.

CHAPTER 7 METHOD FOR THE DESIGN OF A SITUATION AWARENESS GLOBAL ASSESSMENT TECHNIQUE (SAGAT) FOR SIMULATION TRAINING IN ADULT INTENSIVE CARE

7.1 INTRODUCTION

The previous chapter has highlighted some of the challenges associated with the measurement of invisible cognitive skills including SA and other studies have highlighted the difficulty in using observer based tools to measure SA.[153,320,351] Research in military aviation and advanced driving has shown that other techniques of measuring SA such as direct, self-assessment tools and direct objective tools are effective (please see Chapter 2). I decided that a multimodal approach to understanding how SA is best measured would be useful and these next studies will analyse the use of a direct, objective tool (the Situation Awareness Global Assessment Technique: SAGAT[133]); a direct, self-assessment tool (the Situation Awareness Rating Technique: SART[132]) and an indirect, observer based tool (ANTS[152]). This chapter will focus on the extensive background work which must be completed before SAGAT can be used and describe the method of scenario design for teams of healthcare professionals from AICU. SART will be described in more detail in Chapter 9 and ANTS has already been analysed in Chapter 6.

7.1.1 OVERVIEW OF SAGAT

SAGAT is a direct, objective assessment of SA which has been validated in military rescue settings[403], air traffic control[404] , teleoperations [405], military aviators[406] and military operations[407]. More recently in healthcare it has been used to assess SA in trauma physicians and medical students[408]; emergency department physicians[137]; nursing students[409],

multidisciplinary trauma teams[410] and obstetric teams[150]. It has not been used for teams working in intensive care units (ICUs).

SAGAT requires that the goals and sub-goals of any scenario are described. The relevant tasks and decisions for each goal are then determined by subject matter experts to produce a goal-directed task analysis (GDTA). Questions are then designed to be asked of participants at pause points (or "freezes") in a scenario which explore SA relevant to the situation. The questions must be posed to assess the three levels of SA (perception, comprehension and projection). Standardising the language used to ask the questions and ensuring that they are very clear is also vital. This process is quite involved and presents issues for simulation educators in healthcare, which are less evident in the other domains where SAGAT has been used.

7.2 OBJECTIVES

The objectives of this initial phase of the Situation Awareness in Simulation study (SASi) were to:

- Choose two scenarios suitable for use in training teams of multidisciplinary healthcare professionals in AICU, one of which would be more complicated than the other
- Devise a GDTA for both scenarios
- Construct SAGAT questionnaires, which were relevant to all team members, for both scenarios to be used in the SASi study

7.3 CHALLENGES OF DEVELOPING SAGAT FOR HEALTHCARE SETTINGS

The scenarios used for ICU teams in OxSTaR are usually designed with a common high-level structure:

- PHASE 1: An initial relatively stable phase where changes would begin to occur and decisions about necessary investigations or treatment should be made
- PHASE 2: A middle more active phase where the patient is becoming increasingly unwell requiring definitive decisions to made about treatment
- PHASE 3: A final resolution phase where the treatments (if implemented appropriately) take effect and resolve the problem.

In military settings banks of questions are defined by experts that can be used at many points in the scenarios. For example, asking participants to watch a radar screen with a variety of different threats at different altitudes allows similar questions to be asked repeatedly. However, in an ICU scenario there may be some questions which could be asked of any situation but others which require changing to fit the situation. Consider comparing a trauma patient who is 38 weeks pregnant and a trauma patient who is 80 years old. Whilst changes in heart rate and blood pressure will be relevant in both, the implications and treatment will be different and questions will need to be added which are relevant to condition of the baby in the former scenario and relevant to the patient's quality of life and co-morbidities in the second. This means that questions posed at level 1 SA (perception) may be the same (e.g. what is the heart rate?) but those asked at level 2 (comprehension) and level 3 (projection) may need to be phrased differently. Equally importantly, the answers will be different and design of the scoring template for each will require adaptations.

7.4 DEVELOPMENT OF SCENARIOS FOR THE SITUATION AWARENESS IN SIMULATION (SASI) STUDY

A team of 10 subject matter experts was drawn from the adult general ICU (AICU) team in the OUHT and OxSTaR. The experts were two simulation educators and consultant anaesthetists, an

intensive care medicine (ICM) simulation fellow, two senior AICU nurses, two consultants in anaesthetics and ICM, and three simulation technicians. They had a combined experience of over 50 years in the design of simulation scenarios.

Dr Mica Endsley, the original author of the SAGAT tool[80], also provided additional advice on the development of the GDTA.

Two scenarios were chosen and developed for this study of situation awareness in simulation (the SASi study) from a bank of 22 existing scenarios used in ICM and anaesthetics simulation training. The key considerations in choosing them were:

- Did they provide the opportunity to test a range of technical and non-technical skills?
- Did they provide opportunities to test all members of the team across that range of skills?
- Were they presenting sufficiently contextually different stories to reduce learning bias across the two scenarios
- Was one obviously more complicated than the other?

They were adapted iteratively according to methodology presented below and the basic outline was as follows

- Scenario A: A 75 year old patient admitted with sepsis who develops atrial fibrillation requiring DC cardioversion and, subsequently, an asystolic arrest
- Scenario B: A 36 year old multiply-injured patient who arrives in the ICU after a splenectomy with a history of asthma and long Q-T syndrome and goes on to develop a tension pneumothorax followed by ventricular fibrillation.

There were two key differences in the scenarios:

1) Scenario A was deliberately less complicated than scenario B. A comparison of the number of requirements at each level of SA in the GDTA between scenario A and scenario B revealed they were 60% greater in scenario B.
2) In scenario A the junior nurse entered the simulation room first and undertook a series of observations before changes began to occur and in scenario B all team members entered together.

Both scenarios were repeatedly tested in the simulator by six research team members including answering the SAGAT questions to assess time required at each pause.

Both scenarios provided the teams with opportunities to recognise and treat common emergencies encountered on ICUs requiring teams to be aware of changes in the physiological condition of the patient and respond with appropriate treatment in a timely fashion. The scenarios involved familiar clinical situations in ICUs. There is guidance on best practice for managing these emergencies from the UK Resuscitation Council and ICU staff are trained in these protocols.

Both scenarios were designed to have three phases three associated pauses at which questions were asked as described below.

Scenario A:

At the beginning of this scenario the patient is stable and awaiting a pre-arranged CT (computed tomography) scan. The junior nurse enters the room first and has a handover (provided in a standardised way by a faculty member acting as another ICU nurse). The patient then develops intermittent runs of atrial fibrillation (AF) before converting to persistent AF. The situation deteriorates and the team will be required to perform an electrical cardioversion

according to protocol. After cardioversion the patient develops asystole requiring the team to use the appropriate arm of the ALS (Advanced Life Support) treatment algorithm.

Three pause (freeze) points were specified.

- Pause 1: The point at which treatment for the AF is described or initiated (whichever happens first)
- Pause 2: The point at which the definitive decision to perform an electrical cardioversion is made
- Pause 3: The point at which the patient becomes asystolic.

Scenario B:

At the beginning of this scenario the whole team is in the room receiving handover from an anaesthetist (this is provided in a standardised way but a faculty member acting as the anaesthetist from emergency theatre). The patient has a background of childhood asthma and a preoperative ECG which is with the patient's anaesthetic chart reveals long Q-T syndrome (the team is not told explicitly that the patient has this condition). The patient deteriorates after handover and develops initial bronchospasm with desaturation followed by a tension pneumothorax. This requires the team to diagnose and treat the pneumothorax rapidly. The hypoxia caused by the bronchospasm and pneumothorax leads to intermittent ventricular tachycardia followed by ventricular fibrillation. The team is required to use the appropriate arm of the ALS algorithm (shockable rhythms) with prompt provision of a DC shock for defibrillation.

- Pause 1: The point at which treatment for bronchospasm is described or initiated (whichever happens first)
- Pause 2: The point at which the pneumothorax is diagnosed
- Pause 3: The point at which the patient goes into ventricular fibrillation.

7.4.1 PROGRAMMING THE SCENARIOS

Team performance varies considerably during simulation training and faculty in the control room must respond to decisions made in real-time by altering physiological parameters from the computer driving the manikin which mimic what would happen in real life, for example: when a nurse increases the rate of a nor-adrenaline infusion in response to a fall in blood pressure faculty in the control room should increase the blood pressure or if a doctor performs a needle decompression for a pneumothorax faculty should reverse the hypoxia and re-inflate the manikin's lung. The advantage of having programmable manikins is that these responses can be standardised such that blood pressure would return to the same level after an intervention to improve it and the physiological disturbance caused by a pneumothorax would resolve at the same rate after needle decompression in every scenario.

The scenarios were programmed using Laerdal SimDesigner™ software and the modelling of physiological parameters for each is described below. The faculty guidelines, pre-programmed physiological parameters and expected candidate actions for both scenarios are shown in Appendices 22 and 23. The SimDesigner™ programmes and the layout of equipment in the simulation room are shown in Figure 7-1 and Figure 7-2 and Figure 7-3.

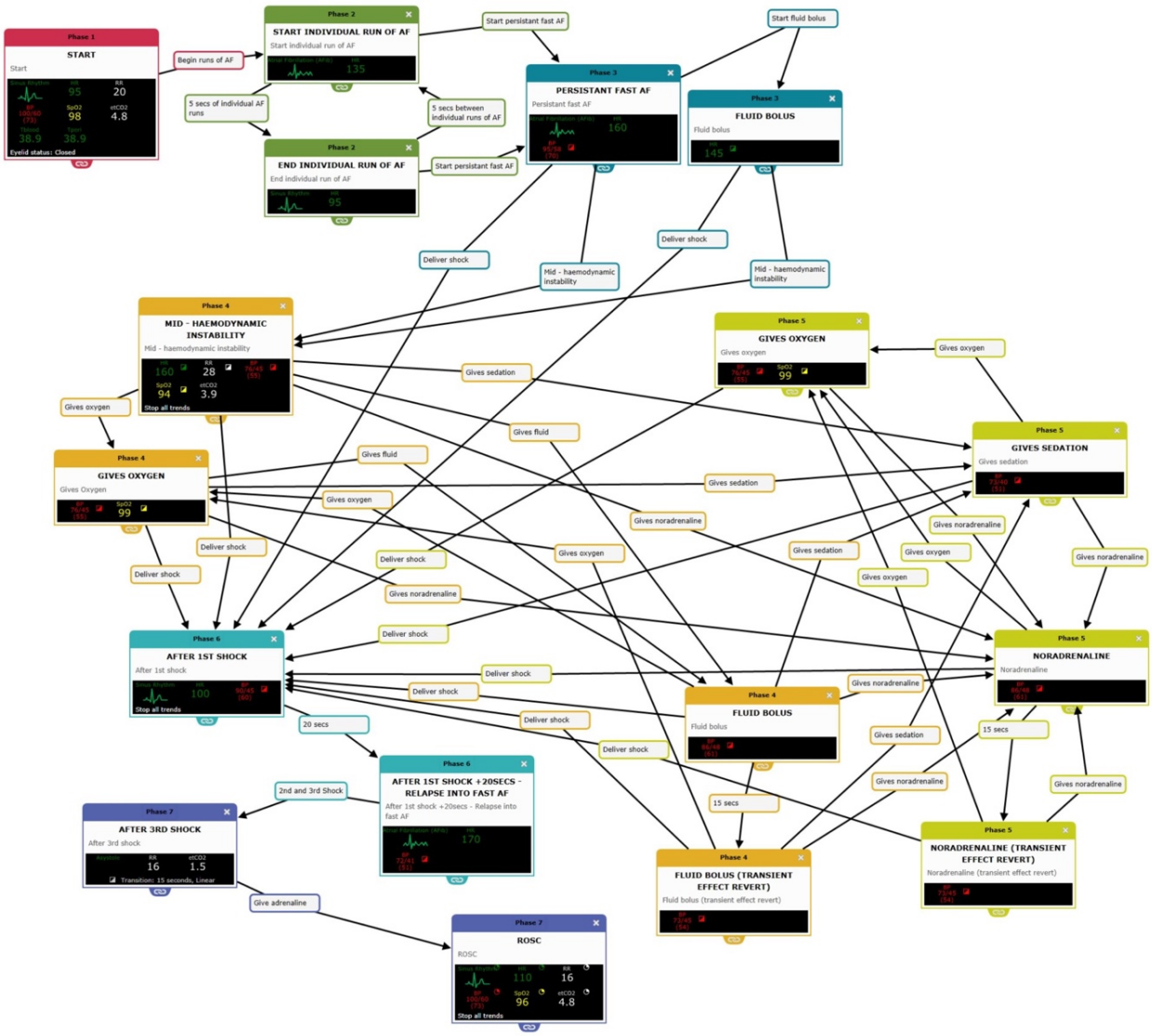


Figure 7-1: SimDesigner™ template for scenario A showing standardisation of response and timing of interventions

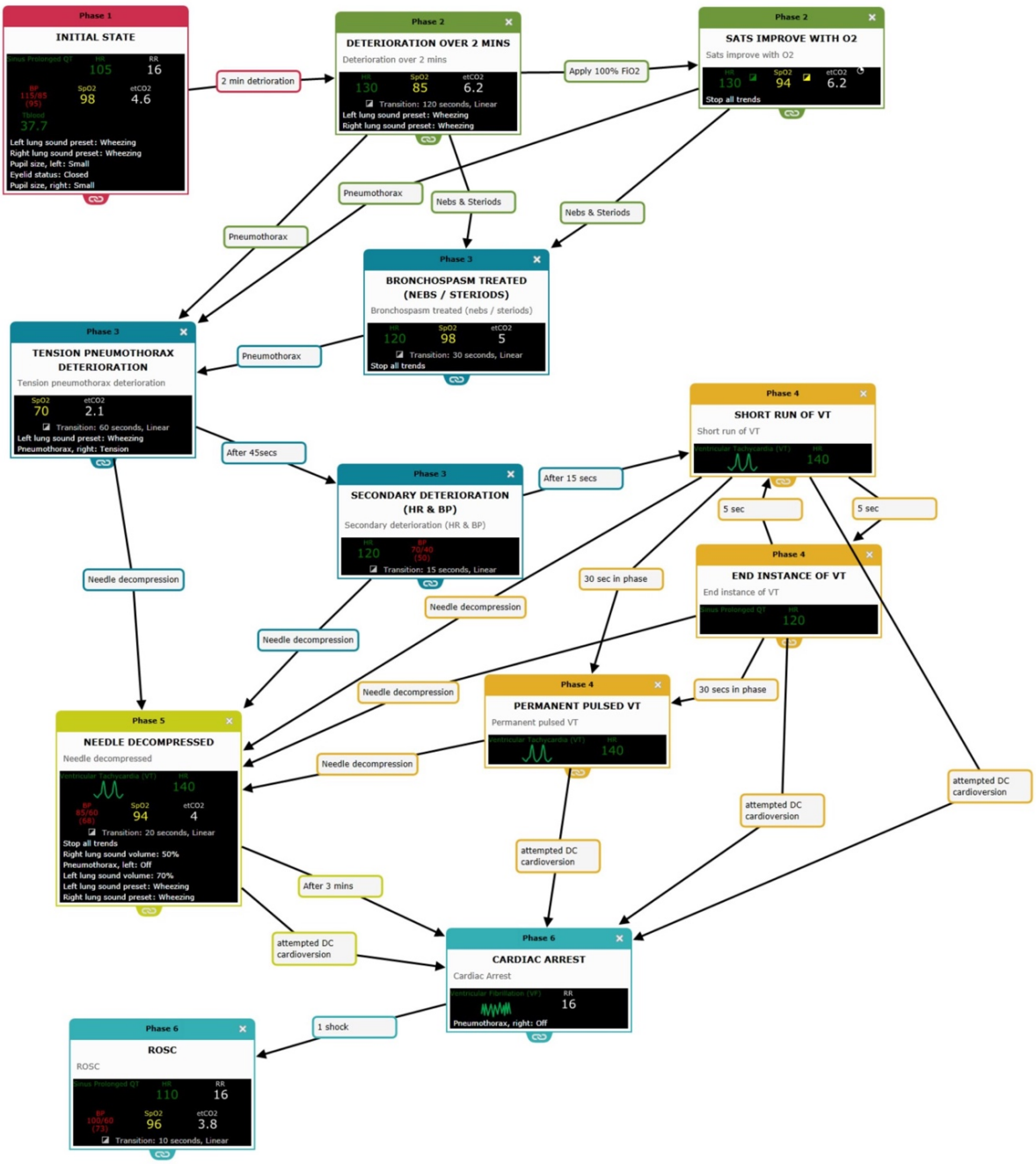


Figure 7-2: SimDesigner™ template for scenario B showing standardisation of response and timing of interventions

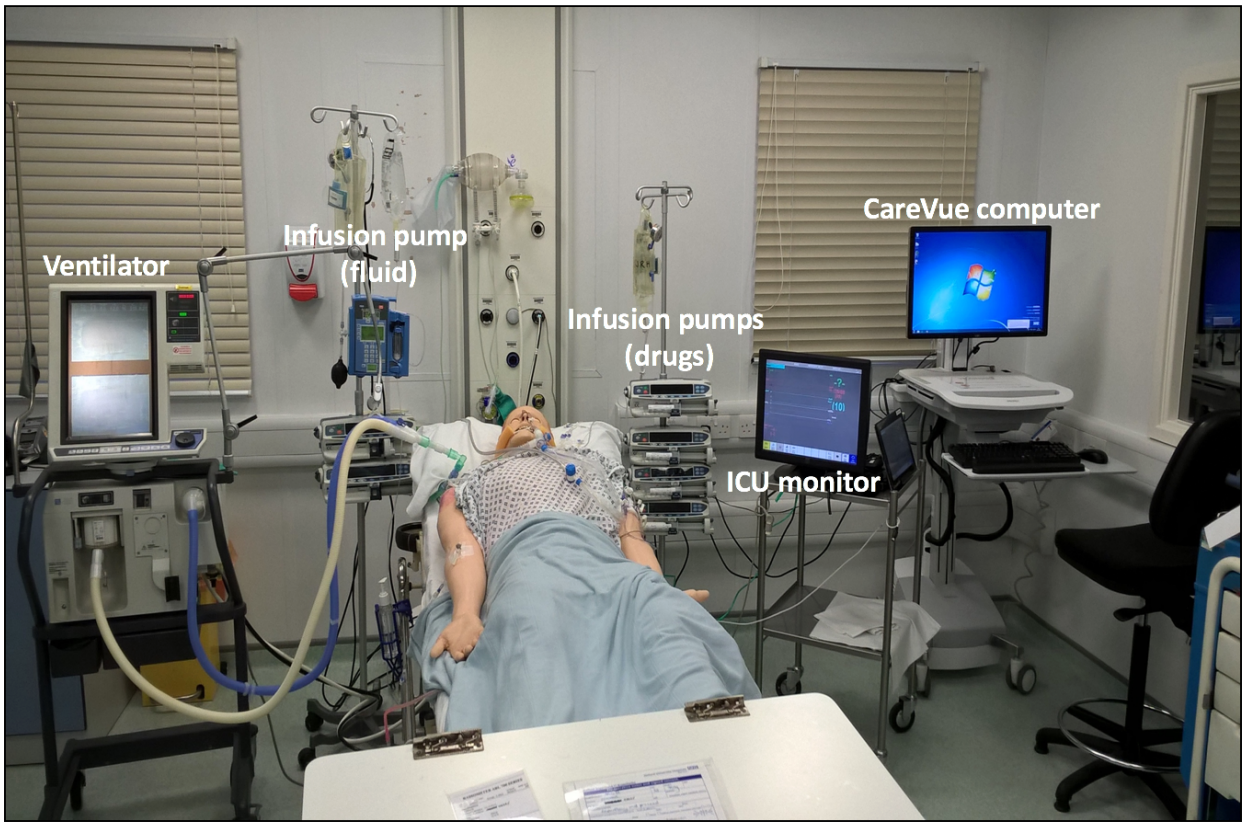


Figure 7-3: Simulation room set up for SASi study using equipment from AICU

7.5 RESULTS: GOAL DIRECTED TASK ANALYSIS (GDTA) AND SAGAT QUESTIONNAIRES

7.5.1 GDTAS

After the scenarios had been chosen the GDTAs were developed in conjunction with the programming of the physiological parameters. The GDTA is a vital component of development of the SAGAT questionnaire. It describes the goals and sub-goals for a given situation which, in turn, inform the situation awareness requirements and then the questions for participants which will explore their SA at all three levels. The GDTAs were developed using a Delphi Method[411] of iterative interrogation and including:

- Context specific expertise of the whole research group

- Task analysis from previous simulation training incorporating these scenarios
- Best practice guidance from the UK Resuscitation Council for the management of atrial fibrillation[412] causing haemodynamic compromise and for cardio-respiratory arrest (Advanced Life Support[382])
- Best practice guidance for the management of sepsis, pneumothorax and trauma from relevant professional bodies (British Thoracic Society[413]; European Society of Cardiology[414]; the National Institute for Health and Care Excellence[415]; the American College of Surgeons Advanced Trauma Life Support course[416])

The final process is shown in Figure 7-4.

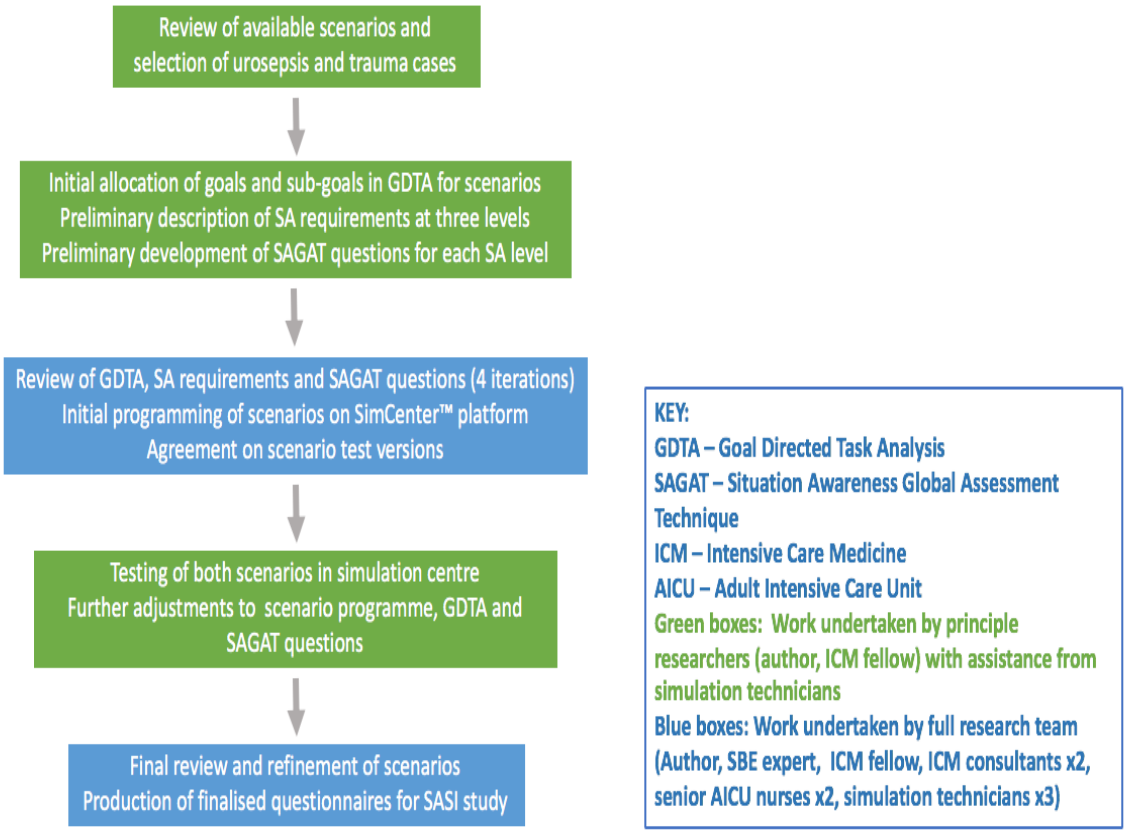


Figure 7-4: Flow diagram of iterative process of design of GDTAs and SAGAT questions for SASi study

The number of questions differed at each point. This was due to the inherent clinical differences within the scenarios and, therefore, relevance of questions regarding the three levels of SA (level 1: perception; level 2: comprehension and level 3: projection) at different time points. Appendices 24 and 25 show the GDTA and questions for both scenarios and Table 7-1 gives examples of questions asked for the three levels of SA at pause two in scenarios A and B.

Scenario A:

- Pause 1: The point at which treatment for the AF is described or initiated (whichever happens first) 12 questions (5 at SA level 1, 5 at SA level 2 and 2 at SA level 3)
- Pause 2: The point at which the definitive decision to cardiovert is made
  6 questions (2 at SA level 1, 2 at SA level 2 and 2 at SA level 3)
- Pause 3: The point at which the patient goes into asystole
  4 questions (1 at SA level 1, 1 at SA level 2 and 2 at SA level 3)

Scenario B:

- Pause 1: The point at which treatment for bronchospasm is described or initiated (whichever happens first)
  10 questions (4 at SA level 1, 4 at SA level 2 and 2 at SA level 3)
- Pause 2: The point at which the pneumothorax is diagnosed
  8 questions (3 at SA level 1, 3 at SA level 2 and 2 at SA level 3)
- Pause 3: The point at which the patient goes into ventricular fibrillation
  4 questions (1 at SA level 1, 1 at SA level 2 and 2 at SA level 3)

Scenario	Questions at pause 2	Answers at pause 2
A	SA1 (perception): What is the patient's heart rate?	± 10% of actual value
	SA2 (comprehension:) What is your main concern at this point?	Either: low blood pressure/cardiac output or failure to respond to treatment instituted if that has happened
	SA3 (projection): What are the next steps?	Two of the following: prepare for cardioversion - set up defibrillator, ensure FiO_2 1.0, ensure adequate sedation; review pharmacological treatments (including electrolyte replacement); check personnel present are competent to deliver cardioversion
B	SA1 (perception): What are the findings on respiratory examination?	Evidence of tension pneumothorax on the right (must get side correct): decreased chest wall movement on right, decreased air entry on right, wheeze on left also audible
	SA2 (comprehension): What is your main concern at this point?	Treating the tension pneumothorax (no score for mentioning VT in isolation)
	SA3 (projection): What are the next steps?	Two of the following: Prepare for immediate decompression of right side of chest (needle or blunt dissection); set up for formal chest drain; prepare for management of deterioration in cardiac output; recognise risk associated with long Q-T and have defibrillator ready; call for help

Table 7-1: Example questions from SAGAT questionnaire for three levels of SA at pause 2 in scenarios A and B in SASi study

7.5.2 ADMINISTERING SAGAT QUESTIONS

Recommendations made by Endsley[147] about the administration of SAGAT and the pauses required include:

- Initial pause should occur after no less than five minutes into the scenario
- Pauses should not last longer than about five minutes (during testing the three pause times for the SAGAT scenarios were all less than three minutes)
- There is no limit to the number of pauses which might occur in a scenario provided that the above guidance on timings are adhered to
- Freeze times should be randomly determined (where scenarios are repeated or very similar)

- Collection of 30-60 samplings per question is desirable
- The simulator computer should be collecting objective data corresponding to questions at the time of the pauses (e.g. accurate blood pressure and heart rate etc)
- There should be no visible data relating to the scenario when participants are answering questions
- Questions asked but not answered by the participants should be considered incorrect
- Questions should be scored as correct or incorrect based on whether the answer falls in an accepted tolerance band
- Care should be taken in combining questions into levels of SA as SA at each query point can be independent of others as a result of shifts in operator attention and differences in task demands in the moment

There were three pauses in both scenarios which aligned with the three phases described above and occurred when team came to pre-determined decisions or actions for each scenario. Pause times were not randomised as the scenarios and, therefore, questions asked were substantially different. The length of each pause was timed and all data from the point at which the pause occurred were stored on the control room computer. The questions were put to the candidates in isolation (they were each taken to different, adjacent rooms: the simulation room; the control room and the seminar room) so that they were not able to use cues from their colleagues or the environment (the monitors were covered) which may have influenced their answers. A member of faculty was present with each candidate whilst they were answering the questions to ensure there were no problems in understanding.

7.5.3 SCORING SAGAT QUESTIONNAIRES

There were 22 questions for each scenario and a scoring template was devised for both alongside the iterative development of the GDTA. All questions regarding physiological parameters were scored as correct if they were within a range of ± 10% of the actual value. Answers which required candidates to use free text required that two or more of a range of answers appropriate to the condition of the patient and to the participant's clinical background be given, for example at pause 2 in scenario B where the patient has developed a tension pneumothorax the question "what are the next steps?" was marked correct where two or more of the following options were described:

1) Call for additional help (doctor trained in ICM or cardiothoracic surgery and more nursing assistance)
2) Recognise need for immediate decompression of right side of chest (doctor to request large bore needle or equipment for blunt dissection, and nurses to gather equipment for chest decompression)
3) Anticipate requirement for a formal chest drain (doctor to request chest drain kit and nurses to gather necessary equipment)
4) Anticipate risks of worsening cardiovascular compromise (doctor to include awareness of long Q-T and risk of "torsades de pointes" arrhythmia, nurses to gather equipment for resuscitation including defibrillator).

Where any free text answers were deemed ambiguous by the primary assessor they were reviewed by additional appropriately qualified members of the team and a consensus decision made.

7.6 DISCUSSION

There is a limited number of studies available using SAGAT in healthcare and none has described the methodology of design of the GDTA and SAGAT questionnaire in detail.[136,137,148,150,409,417–421] This chapter has described the necessary precursor elements to the use of SAGAT in healthcare training. The development of a GDTA and associated SAGAT questionnaires was a lengthy process and required extensive involvement of subject matter experts who worked iteratively over a two month period to deliver two SAGAT scenarios. Evidence of the content validity of the SAGAT tool for use in the SASi study has been provided by:

- In depth review of the iterative development of two standardised scenarios (with one more complex than the other) for teams of ICU staff, to ensure delivery of the same patient responses to interventions or treatments in every case
- Description of the Delphi process (with relevant subject matter experts) for the design of the GDTA
- Explanation of the iterative process of design of the SAGAT questionnaire in alignment with the GDTA

7.7 CONCLUSION

The final two empirical chapters in this thesis will describe the use of these scenarios and their associated SAGAT questionnaires in the SASi study. A particular focus of SASi will be to provide evidence of validity of the tools for measurement of SA and to provide some insight on the relationship of SA with measures of experience, workload and performance.

CHAPTER 8 A STUDY OF SITUATION AWARENESS IN SIMULATION FOR ADULT INTENSIVE CARE (SASI)- A COMPARISON OF SA MEASURES

8.1 INTRODUCTION

This thesis has, so far, focused on analysing SA error in an acute care hospital and techniques for measuring SA in healthcare. The analysis of errors in a series of serious incidents in the OUH found that SA errors occurred in 96% of cases and cognitively effortful errors at level 3 SA (projection) occurred ten times more frequently in high acuity settings such as intensive care. This chapter will describe a study of SA in multidisciplinary Adult Intensive Care Unit (AICU) teams (MDTs) using scenarios simulating acutely deteriorating patients.

Multidisciplinary simulation-based team training (in-situ on the AICU and in OxSTaR) has been a part of the education programme in AICU at the OUHT for the past six years. This type of experiential learning is used to expose both novice and expert staff to routine and unusual emergency clinical situations based on real critical incidents that occurred in AICU. Audio-visual recordings support expert faculty in debriefing teams post-scenario and allow the MDT to experience real-time decision making in a safe, non-judgmental environment where performance of technical and NTS may be reviewed and improved.

SA is one of the NTS that is assessed. Comparison of SA between teams and individuals, and across time, requires a validated measure of SA. Direct, indirect, observer-based and self-assessment methods of measurement of SA have been described in various work settings but more recently interest in assessing SA in healthcare has been growing. However, available tools

have not been properly characterised. The aims of this study were to assess the validity of three SA measurement tools (one direct objective, one direct, subjective (self-assessment) and one indirect, observer-based) in simulated AICU emergency scenarios.

SAGAT is a direct assessment of SA which has been validated in various settings, including healthcare (see previous chapter).

SART[132] is a generic self-assessment tool for recording individual, subjective ratings of SA regarding: demand on subject's resources; supply of subject's resources and understanding of the situation. It was designed for use in military aviation, is easy to use, does not require customisation across different workplaces, and has been shown to correlate with operator performance[422] and workload assessments[423,424]. SART has not been validated in healthcare.

NTS assessment tools for healthcare provide observer-based assessments of SA. The systematic review of the tools available to assess NTS in this thesis revealed 76 different tools (see chapter 4). ANTS[152] measures four NTS categories: teamwork; task management; situation awareness and decision-making. ANTS was chosen as the observer-based assessment of SA in this study because although it is designed for the anaesthetist alone (i.e. the team leader), it measures all three levels of SA with a clear focus on the importance interaction with team members to achieve goals and could, therefore, be considered a surrogate marker of team SA. The categories and elements are appropriate for other specialties (and have been used to design systems of NTS assessment for nurses[267,368]). It was designed to assess NTS in all clinical contexts of an anaesthetist's work (i.e. not just cardiac arrest like OSCAR). The analysis of NTS assessment tools in chapter 5 found inter-rater reliability to be good (ICC = 0.79) and usability to be best.

There is evidence that experience, knowledge and workload have an impact on SA. The relevance of experience and professional background as well as participant assessment of workload can be measured using the NASA-TLX (National Aeronautics and Space Administration – Task -Load Index) questionnaire.[425] The results from this analysis will be covered in Chapter 9.

Objectives and outcome measures for this study (the "SASi" study) are described in Table 8-1:

	Objectives	Analysis of outcomes
1 (primary objective)	To investigate the construct validity (content, discriminant and relationship with other measures) of SAGAT, SART and ANTS in measuring SA in multidisciplinary staff from the Adult Intensive Care Unit managing simulated emergency situations.	Design of Goal Directed Task Analysis for SAGAT. Separation of SAGAT, SART and ANTS scores between two different scenarios. Correlation of SAGAT, SART and ANTS scores.
2	To assess whether SAGAT and SART perform differently in individuals or professional groups.	Separation of SAGAT and SART scores between individual participants and teams.
3	To analyse the discriminatory capacity of SAGAT at different levels of SA in two AICU simulation scenarios with different levels of SA required.	Separation of SAGAT scores at different levels of SA.
4	To analyse the discriminatory capacity of SAGAT at different time points in two AICU simulation scenarios, where the time points require differing levels of SA.	Separation of SAGAT scores at different time points in two AICU simulation scenarios.
5	To analyse the usability of SAGAT and SART	Quantitative and qualitative measures of usability including time taken to construct SAGAT questionnaires, number of correctly completed SART forms and feedback from participants and facilitators using both.

Table 8-1: Primary and secondary objectives (with analysis of outcomes) to assess validity and reliability of SA measures for SASi study

Ultimately, the aim is to use the information acquired to decide the most effective way of measuring SA in AICU and inform the design and delivery of effective SA-focused simulation based teaching programmes to improve team performance in a crisis.

8.2 METHODS

8.2.1 DESIGN

This study was an observational study of both real-time performance and post hoc review of recordings of AICU MDTs in simulated emergency scenarios. Three different tools for the measurement of SA were used for this part of the study (as described above). The study used mixed quantitative and qualitative methods for the measurement of the demographic profile of participants, situation awareness, and performance. Measures of experience, technical performance and workload were also recorded but they will be described in the next chapter.

This study was submitted to the CTRG at the University of Oxford and deemed to be exempt from ethics review as it forms a part of the ongoing quality assessment of the training programme for multidisciplinary teams in AICU.

8.2.2 PARTICIPANTS

Three team members (one doctor and two nurses) were allocated to each session and, where possible, they included a senior and a junior nurse. Allocation of staff to attend the training was decided on the day according to clinical requirements on AICU. The education and management team on AICU were closely involved in the decision making to ensure that the training would not impinge on clinical care of patients. It was decided that a maximum of three hours per session could be allocated. Trainees in intensive care medicine (ICM) were invited to attend training when their shifts coincided with the sessions. All staff attended training between 09:00 and 17:00 and the training was not offered to staff who had just been on night shifts or on-call.

All staff were asked to provide consent for recording of scenarios in OxSTaR during their simulation training and the standard consent form in the centre includes a provision for storing the recordings for up to one year after the end of training. The consent form is included as Appendix 26.The course feedback from is included as Appendix 27.

Participants in the study were allocated a unique index number and their data (including video files) were stored securely on a password protected University computer in a locked office (in line with the University of Oxford's best practice guidance for researchers). Data were collected from each participant on professional background and number of years of working on AICU.

8.2.3 PROCEDURES – SCENARIOS AND TRAINING SESSIONS

Two pre-existing scenarios depicting common AICU emergencies were standardised for this study and GDTAs with associated SAGAT questions generated as described in the previous chapter. The testing process clarified the additional time required to include SAGAT and it was clear that it could only be applied to one scenario per session in order to comply with overall time allotted for the training (three hours). We therefore pre-randomised the scenarios for each training session both by order and by inclusion of SAGAT questionnaire to reduce the impact on the results of observer and operator bias, risk of participant learning between scenarios and the impact of one scenario being intrinsically more complicated than the other.

Additional data gathered from each participant were: SAGAT questionnaires and SART questionnaires (see below). NASA TLX scores were also completed and these will be described in the next chapter.

8.2.3.1 THE SIMULATION TRAINING SESSION

The study scenarios were run in the simulation centre (as opposed to the AICU) in order to standardise the surroundings in which the SA assessments were taking place.

The simulation suite in OxSTaR is comprised of three adjacent rooms: a control room from where faculty and technical staff run the scenario and observe the team's performance, the simulation room (in the middle) where the action takes place and the seminar room where the debriefing is conducted after the scenario. The simulation room has one-way mirrors in the two side walls to allow viewing of the scenario from the control room and the seminar room. OxSTaR houses a variety of hi-fidelity, wireless-controlled adult patient manikins and audio-visual recording equipment.

Teams of three participants arrived at OxSTaR shortly before the training session was due to begin and completed a pre-training questionnaire regarding their confidence levels in emergency situations and the confidentiality agreement. They were asked not to share details of the scenarios used in the training to avoid biasing the responses of future participants.

The teams comprised one doctor and two nurses. Where possible, within the constraints of clinical requirements on AICU, one nurse was more senior than the other. All participants took part during normal working hours (i.e. 09:00-17:00). All team members were advised to regard the scenario as a real situation and to treat the simulated patient in the way that they would in real-life (i.e. they were not asked to behave as if they were in another role).

The equipment (ventilator, infusion pumps etc) in the simulation room were the same as those used in AICU and the simulated physiological monitor was formatted to simulate the standard layout on AICU. The electronic patient record system (Phillips ICIP) was pre-filled with data for

the simulated patients. The team members could interact with all of this equipment in the same way that they would in real life. Relevant test results and patient data including anaesthetic charts, chest X-rays, ECGs and arterial blood gases (ABGs) were provided as requested and the same results were shown to all teams involved in the study.

Participants were given the opportunity to ask questions prior to the session beginning. The faculty (one of the investigators or another clinically qualified member of the OxSTaR team) gave a short, standardised briefing on the scenario. A faculty member was present in the room for each scenario in the role of a healthcare assistant. Table 8-5 shows one of the teams in action (consent was given to use the photo).

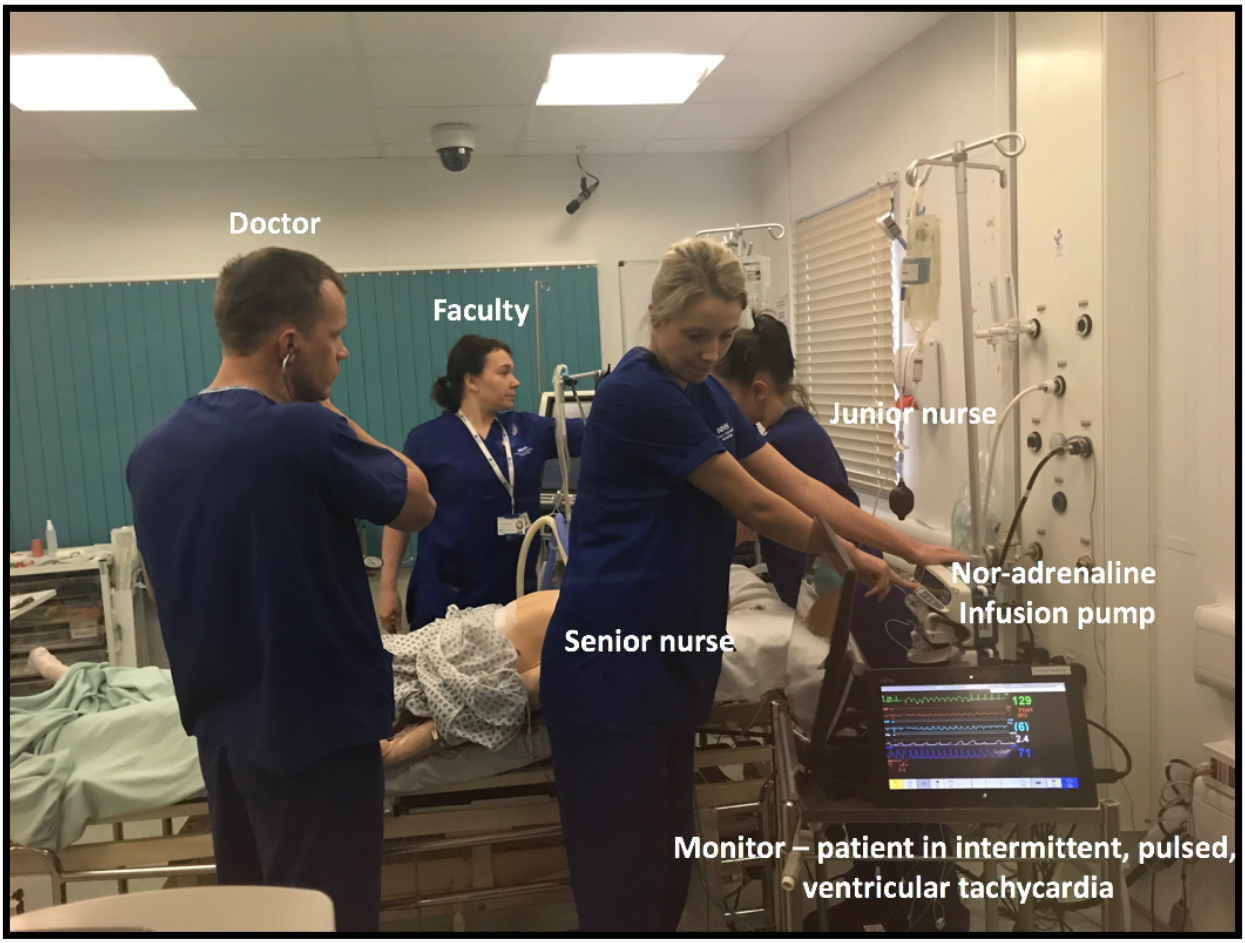


Figure 8-1: Team involved in SASi scenario B. The monitor shows a change in the patient's heart rhythm which the senior nurse is observing but the doctor is unaware of.

The two scenarios (A [easier] and B [more complex]) were run consecutively in the morning or afternoon (according to the randomly assigned order). Differences in pause points and SA requirements for SAGAT in each scenario are described in the previous chapter. Debriefing was completed immediately after the scenarios by two experienced faculty members (HH and LV) using the video recordings to support objective feedback with a particular focus on the NTS (including SA) in the team performance.

8.3 MEASURES

8.3.1 MEASUREMENT OF SA

SA was measured with SAGAT,[133] SART,[132] and ANTS,[152] tools as described above.

8.3.1.1 SAGAT

SAGAT questions were designed and used as described in the previous chapter.

8.3.1.2 SART

The SART[132] tool originated in military aviation and scores are generated by the participants answering 10 questions spread across three domains (10-D SART) on a seven point, low-to-high, Likert scale. The three domains are: demand (three questions), supply (four questions) and understanding (three questions). These domains may be considered separately or as an overall score by calculation of mean values for each domain and use of the following formula: SA(calc)= Mean understanding – (mean supply – mean demand).[426] However, because of the lack of reliability evident in this manipulation of SART scores[426] I decided to use the original 10-D scale

and analyse the means of the three domains of demand, supply and understanding separately.[134]

A standardised explanation was provided of the construct of the questions and how to answer them by the author at the start of each training session. Participants answered the questionnaires, using the seven point Likert scale, after each scenario (the SART questionnaire is included as Appendix 28).

8.3.1.3 ANTS

ANTS assessments were made by HH using video review of the 38 scenarios after completion of the study. The use of the ANTS system has been described extensively in Chapter 6.

8.3.2 MEASURES OF EXPERIENCE, PERFORMANCE AND WORKLOAD

Measures of experience (number of years working in ICU), technical performance (using an observer based global performance score and time taken to achieve complete the scenario) and workload (using the NASA-TLX system) were recorded for each team. The results of these assessments will be discussed in the Chapter 9, but are included in Table 8-2 for completeness.

Timing of measurement	Measurement tool
Measurements before scenario	Feedback questionnaire: pre-training questions for 57 participants (including participant experience level)
Measurements during scenario	**SAGAT questionnaire** for 19 scenarios (used in each session for one scenario). NASA-TLX for 19 scenarios (to be discussed in the following chapter). Observer based assessments of NTS including SA using ANTS to *inform the debrief* for all scenarios (not formally scored).
Measurements after scenario	**SART questionnaire** for 57 participants (both scenarios). Feedback questionnaires: post-training questions and free text for 57 participants.
Measurements after study completion	Observer-based assessment of global technical performance for teams in 38 scenarios (HH and LV). **Observer-based assessment of NTS** including SA using ANTS for 38 scenarios (HH). Measurement of total action period for 38 scenarios (HH). Scoring of SAGAT questionnaires (HH).

Table 8-2: Time of use of different measurement tools in SASi. Measurements explored in this chapter are highlighted in bold script, the measures of experience, workload and performance will be described in Chapter 9.

8.3.3 USABILITY

The usability of SAGAT and SART was assessed for participants and faculty by:

Participants:

- measuring completeness of questionnaires
- time taken to complete questionnaires (SAGAT only)
- analysis of feedback forms

Faculty:

- time taken to develop questionnaires (SAGAT only)
- time taken to analyse questionnaires
- qualitative feedback from subject matter experts involved in delivering the training

The usability of ANTS has been described in chapter 6. -

8.4 DATA ANALYSIS

SA scores were recorded and means and percentages were calculated as appropriate for all numeric variables. SAGAT answers were analysed by measuring the accuracy of individual scores (summed scores with percent correct). Assessments of normality of distribution of data were made for the SA measures and parametric and non-parametric measures of difference and correlation used accordingly. All recorded data were used in the analyses. The analyses for ANTS and SART were for all 19 teams, but SAGAT scores were only available for the eight teams in scenario A and for 11 in scenario B.

8.4.1 ANALYSIS OF VALIDITY

Evidence of validity was analysed for each SA measurement tool under the following domains:

- **Content validity** was measured qualitatively.
- **Discriminant validity** was analysed by measuring differences between scenarios, professional groups (doctors, senior nurses and junior nurses and paired combinations thereof), teams (measured as whole teams).

  Independent group t-tests were used for dichotomous variables and assessment scores and repeated measures t-tests for differences between participant scores between scenarios.

  One-way ANOVA was used to assess inter-professional differences within scenarios and repeated measures (RM) ANOVA was used for comparison of SAGAT scores across pause points and SA levels for different professional subgroups.
- **Relation to other measures** were assessed by exploring correlations between the SA measurement tools using Pearson product-moment correlation coefficients for normally

distributed variables and Spearman's rank-order correlation for non-normally distributed variables.

8.4.2 ANALYSIS OF RELIABILITY

Different measures of reliability of SAGAT SART and ANTS were required:

- Agreement between participants by professional group in each scenario for SAGAT (using Fleiss's kappa, a variant of Cohen's kappa for more than one rater) as described by Morgan et al[150]
- Internal consistency, using Cronbach's alpha for SART and ANTS

This is because SAGAT answers are measured with a binary score (correct or incorrect) and Cronbach's alpha requires an ordinal or continuous scale.

8.4.3 ANALYSIS OF USABILITY

Usability of SAGAT and SART were assessed using quantitative (e.g. number of missing data points, time of SAGAT pauses) and qualitative (e.g. free text questionnaires for participants) methods.

8.5 RESULTS

The study took place between September 2017 and April 2018 with 57 staff (19 teams) participating, 33% male and 67% female (19 doctors [63% male, 37% female] and 38 nurses [18% male, 82% female]). Twenty-five sessions were booked, five were abandoned due to pressure of clinical work, and there was a computer failure in one which prevented the scenarios from running in a standardised fashion. It was not possible for consultants to attend

the training as they were providing clinical cover on AICU for the trainees. SAGAT questionnaires were applied to scenario A on eight occasions and scenario B on eleven occasions. The pause points were determined by team decisions or actions as described in Chapter 7 (all pauses were timed).

This section will now consider evidence for the validity (content, discriminant and concurrent), reliability and usability of SAGAT, SART and ANTS. There are more data for SAGAT in this chapter than for SART or ANTS partly because SAGAT required extensive input in design to ensure content validity and, because SART was found to have poor reliability in this study (see below) and ANTS has been described in detail in Chapter 6.

8.5.1 EVIDENCE FOR CONTENT VALIDITY OF SA MEASURES

8.5.1.1 SAGAT

The iterative process of designing and refining the two scenarios and developing a GDTA and SAGAT questions for each provided robust evidence of face and content validity (including additional advice from the author of the SAGAT tool, Dr Mica Endsley) and is described in Chapter 7.

8.5.1.2 SART

SART was designed as a generic tool for the subjective assessment of SA in individuals. The questionnaire was reviewed by the same clinical staff (intensivists and nurses) and educators involved in the development of the GDTA for SAGAT (see Chapter 7). It was agreed by all that the questions were appropriate for the scenarios but the review highlighted some concerns which were summarised from verbal feedback:

- the “demand” and “supply” domains appeared more focused on workload than SA
- the Likert scale was 1-7, “low to high” but for one of the questions there was an indirect relationship to the construct described: question 4 in the supply domain asks, “what was the degree to which your attention was divided in the situation?”, if the answer were “high” this would suggest a reduction in supply of attentional resources. All other questions in this domain had a direct relationship with supply of attentional resources.

8.5.1.3 ANTS

ANTS has been found to have excellent face and content validity through the rigorous design processes of the original authors and examination by others using it for assessment of NTS[152,267,312,368] (including Chapter 6 of this thesis).

8.5.2 EVIDENCE OF DISCRIMINANT VALIDITY OF SA MEASURES

8.5.2.1 ANALYSIS OF DIFFERENCE IN ACCURACY OF SAGAT SCORES BETWEEN SCENARIOS

The SAGAT questions for each participant were marked either correct or incorrect according the score sheet derived from the GDTA as described previously. All questions regarding physiological variables were allowed an error range of ±10%. Where no answer was given the score allocated was “incorrect”. The marking of the SAGAT score sheets took approximately 60 hours (i.e. one hour per questionnaire). Additional time was taken to ensure that any ambiguity of answers was agreed with the appropriate context specific expert. This happened infrequently because the answer template included most of the possible and acceptable

answers for doctors, senior nurses and junior nurses and the questions had been designed to be generalisable to all participants (see Chapter 7).

Questions explored three levels of SA in both scenarios: 8 questions for level 1 SA; 8 questions for level 2 SA and 6 for level 3 SA, a total of 22 questions per scenario.

The number of questions at each pause point was dependent on the circumstances and differed between scenarios (see previous chapter). Results were calculated as percentage correct answers to give an "accuracy score" for SA levels and for pause points for the whole team, sub-teams and individual professional groups. These results are shown in Table 8-3, Figure 8-2 and Figure 8-3.

Professional group	Scenario A			Scenario B		
Overall Accuracy						
Whole team	70% (370/528)			62% (452/726)		
Senior team	69% (242/352)			44% (318/726)		
Senior and junior nurse	72% (254/352)			39% (282/726)		
Doctor and junior nurse	69% (244/352)			42% (304/726)		
Doctor	66% (116/176)			70% (170/242)		
Senior nurse	72% (126/176)			61% (148/242)		
Junior nurse	73% (128/176)			55% (134/242)		
Accuracy at SA levels						
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Whole team	65% (125/192)	61% (117/192)	89% (128/144)	48% (128/264)	58% (152/264)	87% (172/198)
Senior team	63% (80/128)	60% (77/128)	89% (85/96)	54% (95/176)	61% (107/176)	88% (116/132)
Senior and junior nurse	66% (85/128)	64% (82/128)	91% (87/96)	44% (77/176)	51% (90/176)	87% (115/132)
Doctor and junior nurse	66% (85/128)	59% (75/128)	88% (84/96)	48% (84/176)	61% (107/176)	86% (113/132)
Doctor	63% (40/64)	55% (35/64)	85% (41/48)	58% (51/88)	70% (62/88)	86% (57/66)
Senior nurse	63% (40/64)	66% (42/64)	92% (44/48)	50% (44/88)	51% (45/88)	89% (59/66)
Junior nurse	70% (45/64)	63% (40/64)	90% (43/48)	38% (33/88)	51% (45/88)	85% (56/66)
Accuracy at Pause Points						
	P1	P2	P3	P1	P2	P3
Whole team	54% (154/288)	90% (129/144)	91% (87/96)	50% (166/330)	71% (186/264)	76% (100/132)
Senior team	50% (95/192)	94% (90/96)	89% (57/64)	54% (119/220)	74% (131/176)	77% (68/88)
Senior and junior nurse	58% (112/192)	90% (86/96)	88% (56/64)	44% (97/220)	66% (116/176)	78% (69/88)
Doctor and junior nurse	53% (101/192)	85% (82/96)	95% (61/64)	53% (116/220)	71% (125/176)	72% (63/88)
Doctor	44% (42/96)	90% (43/48)	97% (31/32)	63% (69/110)	80% (70/88)	71% (31/44)
Senior nurse	55% (53/96)	98% (47/48)	81% (26/32)	45% (50/110)	69% (61/88)	84% (37/44)
Junior nurse	62% (59/96)	81% (39/48)	94% (30/32)	43% (47/110)	63% (55/88)	73% (32/44)

Table 8-3: SAGAT scores (percent correct with actual scores and relevant totals in parentheses) for all participants by professional group, teams and pairings

Independent measures t-tests were used to analyse the difference in SAGAT scores between each scenario. Tests of normality for all SAGAT scores revealed two outliers as assessed by inspection of boxplots (>1.5 SD). Inspection of their values did not reveal them to be extreme and they were kept in the analysis. SAGAT scores from scenario A were not normally distributed

(Shapiro Wilk <0.05), but were normally distributed for scenario B (p > 0.05). The difference between total SAGAT scores for all participants was analysed using an independent samples t-test with interpretation of results using Welch's method where homogeneity of variance was violated (i.e. Levene's test for equality of variances: p=<0.05).

Mean SAGAT scores for all participants (not grouped into teams) were significantly higher in scenario A (70.1% ±7.3), which was the easier scenario, than scenario B (62.3 ± 14.4), which was more complex, a difference of 7.8% (95% CI, 2.0 to 14.0), p = 0.01. This provides evidence that the SAGAT measure was able to discriminate between the two scenarios.

When SAGAT scores were analysed for professional groups doctors were found to have higher scores in scenario B (70.2% ± 12.3) (the more complex scenario) than scenario A (65.9% ± 8.1) (the easier scenario) but this situation was reversed for the nurses: senior nurses scored more highly in scenario A (71.6% ± 8.3) than scenario B (61.2% ± 14.6) and the same was true for junior nurses (scenario A [72.7% ± 3.4] and scenario B [55.4% ±13.3]). The difference was only statistically significant for the junior nurses: p = 0.001 where a mean difference of 17% equated to 4 correct answers' difference. This provides evidence that the SAGAT tool was able to discriminate between professional groups – the junior nurses entered the simulation room before any other team members and, therefore had more time to acquire SA than their colleagues but all team members began scenario B together.

8.5.2.2 ANALYSIS OF DIFFERENCE IN ACCURACY OF SAGAT SCORES FOR PROFESSIONAL GROUPS AT EACH LEVEL OF SA

A repeated measures ANOVA was used to analyse the difference in accuracy of SAGAT scores (percent correct) within professional groups and whole teams in scenarios A and B. Mauchly's

test of sphericity was not violated for any data. The Bonferroni correction for the three pairwise comparisons (level 1 SA with level 2 SA; level 2 SA with level 3 SA and level 1 SA with level 3 SA) gave a threshold for statistical significance of p=0.017. Results are presented in Figure 8-2 and Table 8-4.

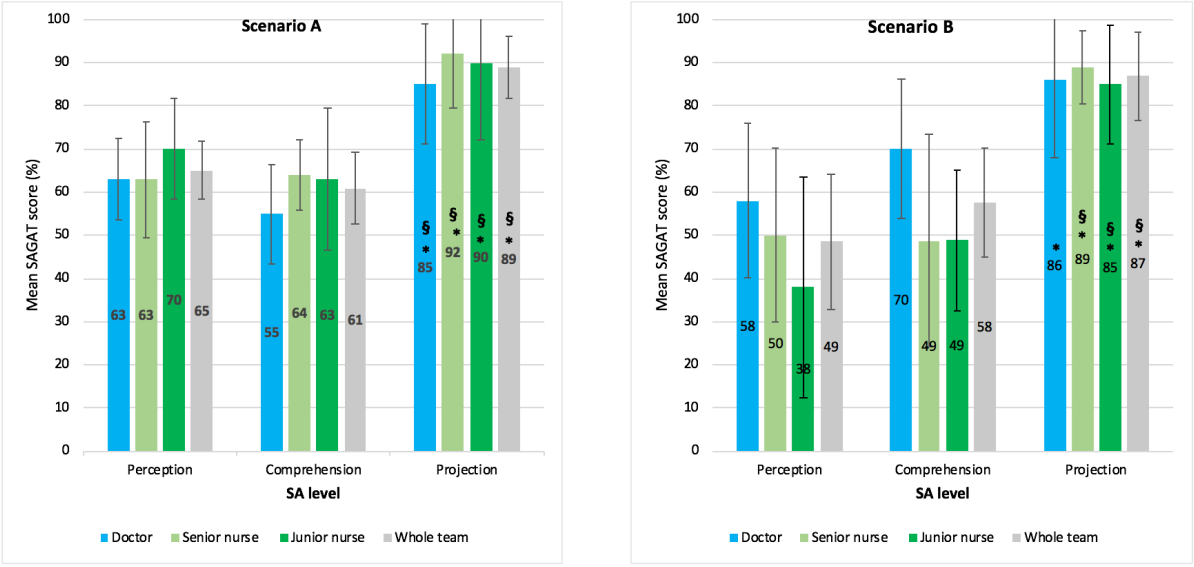


Figure 8-2: Mean SAGAT accuracy scores (%) with SD at 3 SA levels (perception, comprehension, projection) in scenarios A (the easier scenario) and B (the more complex scenario). RM ANOVA revealed significantly higher scores at SA level 3: projection versus level 1: perception (*p < 0.01) and level 3: projection versus 2: comprehension (§p= < 0.01)

SCENARIO A	SA1-SA2		SA2-SA3		SA1-SA3	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
Doctor	7.8 (-14.9,30.5)	0.65	-30.7 (-45.3,-16.2)	**<0.0005**	-21.9 (-43.8,-2.1)	**0.01**
Senior nurse	1.6 (-13.1,10.0)	1.0	-27.6 (-47.1,-8.1)	**0.009**	-29.2 (-45.5,-12.8)	**0.003**
Junior nurse	1.1 (-20.7, 23.0)	1.0	-40.5 (-62.8,-18.2)	**0.001**	-39.4 (-56.5,-22.2)	**0.001**
SCENARIO B	**SA1-SA2**		**SA2-SA3**		**SA1-SA3**	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
Doctor	-11.6 (-27.7,5.0)	0.22	-17.0 (-36.5, 2.4)	0.09	-28.4 (-46.7,-10.0)	**0.004**
Senior nurse	1.1 (-20.7, 23.0)	1.0	-40.5 (-62.8, -18.2)	**0.001**	-39.4 (5-6.5, -22.2)	**<0.0005**
Junior nurse	-7.4 (-35.8, 13.0)	0.63	-36.0 (-56.8, -15.2)	**0.002**	-47.3 (-66.2, -28.5)	**<0.0005**

Table 8-4: Results of repeated measures ANOVA for mean SAGAT scores between SA levels within professional groups), significant results are highlighted in red

Accuracy scores for level 3 SA (projection) were significantly higher than for level 1 (perception) and level 2 (comprehension) for all professional groups in both scenarios (the increase in score between level 2 and 3 for doctors in scenario B did not reach significance). It is interesting that level 3 (projection) scores were higher than level 1 (perception) as it may be assumed that good level 1 SA is a prerequisite for good level 3 SA. It may be that the questions asked at level 1 in the SAGAT questionnaire required too much detail – this will be explored later in this chapter.

8.5.2.3 ANALYSIS OF DIFFERENCE IN ACCURACY OF SAGAT SCORES FOR PROFESSIONAL GROUPS AT EACH PAUSE POINT

A repeated measures ANOVA was used to analyse the difference in accuracy of SAGAT scores (percent correct) within professional groups and whole teams in scenarios A and B. Where Mauchly's test of sphericity was violated the Greenhouse and Geisser correction[427] was applied. The Bonferroni correction for the three pairwise comparisons (pause 1 with pause 2; pause 2 with pause 3 and pause 1 with pause 3) gave a threshold for statistical significance of p=0.017. Results are presented in Figure 8-3 and Table 8-5.

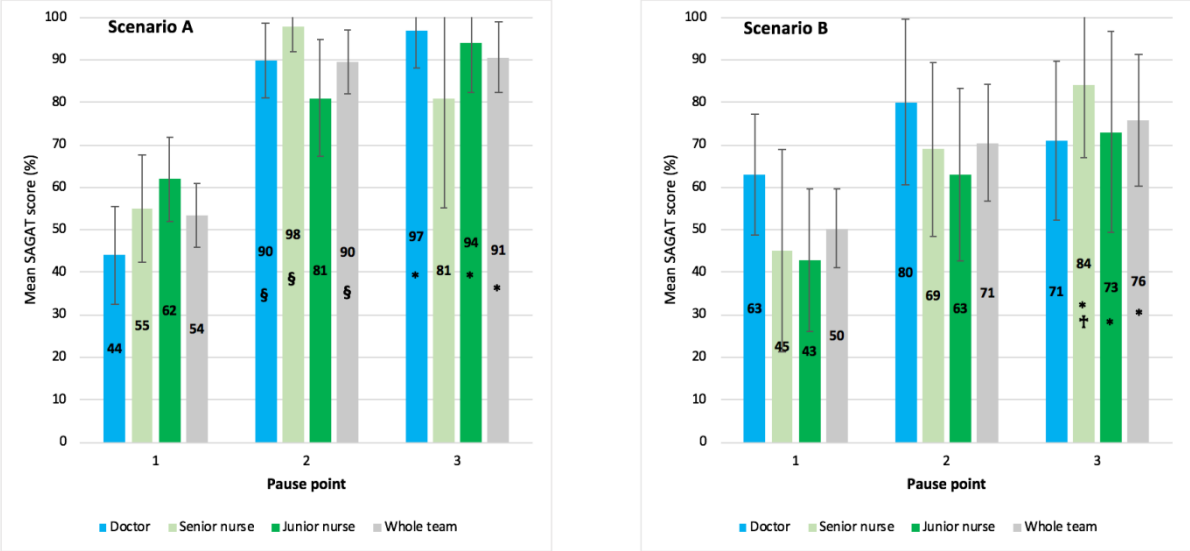


Figure 8-3: Mean SAGAT accuracy scores (%) with SD at pause points in scenarios A and B. RM ANOVA revealed significantly higher scores at pause 3 versus 1 (*p < 0.01) pause 3 versus 2 (†p< 0.01), and pause 2 versus 1 (§p < 0.001)

SCENARIO A	P1-P2		P2-P3		P1-P3	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
Doctor	-45.6 (-61.1,-30.2)	**<0.0005**	-7.5 (-19.1,4.1)	0.248	-53.1 (-62.8,-43.5)	**<0.0005**
Senior nurse	-42.6 (-60.3,-25.0)	**<0.0005**	16.6 (-14.5, 47.8)	0.4	-26.0 (-56.0,4.0)	0.09
Junior nurse	-20.0 (-43.8, 4.1)	0.1	-12.5 (-33.3, 8.3)	0.3	-32.4 (-51.7,-3.0)	**0.004**
SCENARIO B	**P1-P2**		**P2-P3**		**P1-P3**	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
Doctor*	-17.1 (-35.6,1.4)	0.07	9.4 (-8.7,27.5)	0.5	-7.7 (-26.6,11.2)	0.8
Senior* nurse	-24.1 (-51.8, 3.6)	0.1	-14.5 (-25.1,-4.0)	**0.008**	-38.6 (-64.0,-13.3)	**0.004**
Junior nurse	-20.0(-42.4,2.4)	0.08	-10.0 (-31.0, 11.0)	0.6	-30.0 (-51.6, -8.4)	**0.008**

Table 8-5: Results of repeated measures ANOVA for mean SAGAT scores between pause points within professional groups (*Greenhouse and Geisser correction applied), significant results are highlighted in red.

Whole team SAGAT scores increased with time in both scenarios, but there were some underlying differences in the professional groups. In scenario A the difference in SAGAT scores at the first pause was significant for the doctor and the senior nurse but not for the junior nurse. The junior nurse was the first person in the simulation room for this scenario and, therefore, had more time to gather and interpret information early on. In scenario B the doctor

(team leader) was interacting with both nurses from the beginning and all staff heard the handover together.

8.5.2.4 ANALYSIS OF DIFFERENCE IN ACCURACY OF SAGAT SCORES BETWEEN PROFESSIONAL GROUPS AT EACH LEVEL OF SA

A one-way (between-subjects) ANOVA test was used to compare SAGAT scores for each professional group within scenario A and scenario B at the three SA levels within each scenario. Tests of normality found that data were normally distributed for each group (Shapiro-Wilk > 0.05) and Levene's test for homogeneity of variances was not violated for any data set. Tukey's post-hoc test was used to analyse pairwise differences (i.e. between the doctor and the senior nurse, the senior nurse and the junior nurse and the doctor and the junior nurse). The results are shown in Table 8-6.

SCENARIO A	Dr and SN		SN and JN		Dr and JN	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
SA1	0.00 (-14.6,14.6)	1.0	-6.3 (-20.8,8.3)	0.54	-6.3 (-20.8,8.3)	0.54
SA2	-9.4 (-25.0,6.3)	0.31	1.6 (-14.1,17.2)	0.96	-7.8 (-23.5,7.8)	0.43
SA3	-6.3 (-25.0, 12.5))	0.68	2.1 (-20.8,16.7)	0.96	-4.2 (-23.0,14.6)	0.84
SCENARIO B	Dr and SN		SN and JN		Dr and JN	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
SA1	8.0 (-14.6,30.5)	0.66	12.5 (-10.1,35.1)	0.37	20.5 (-2.1,43.0)	0.08
SA2	20.5 (0.02,40.9)	**0.05**	0.0 (-20.4,20.4)	1.0	20.5 (0.02,40.9)	**0.05**
SA3	-3.0 (-17.7,11.7)	0.87	4.5 (-10.2,19.2)	0.73	1.5 (-13.1,16.2)	0.97

Table 8-6: One-way ANOVA for comparison of mean SAGAT scores between professional groups in scenario A and B), significant results are highlighted in red. SA1 = SA level 1 (projection), SA2= SA level 2 (comprehension), SA3 = SA level 3 (projection); Dr = doctor, SN = senior nurse, JN = junior nurse.

There were no significant differences between team members at different levels of SA in scenario A. However, the doctors in scenario B had significantly greater accuracy scores at level 2 SA (comprehension) than the nurses. It is possible that this indicates inadequate communication of the situation by the doctor to the nurses.

8.5.2.5 ANALYSIS OF DIFFERENCE IN ACCURACY OF SAGAT SCORES BETWEEN PROFESSIONAL GROUPS AT EACH PAUSE POINT

A one-way (between-subjects) ANOVA test was used to compare SAGAT scores for each professional group within scenario A and scenario B at each pause point. Tests of normality found that data were normally distributed for each group (Shapiro-Wilk > 0.05). Where Levene's test for homogeneity of variance was violated Welch's ANOVA was reported and Games-Howell's post hoc test used to analyse pairwise differences (this only happened at pause 2 in scenario A for the senior and junior nurse). The results are shown in Table 8-7.

SCENARIO A	Dr and SN		SN and JN		Dr and JN	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
Pause 1	-11.5 (-25.9, 2.9)	0.13	-6.1 (-20.5, 8.2)	0.54	-17.6 (-32.0, -3.2)	**0.02**
Pause 2*	-8.5 (-18.5, 1.5)	0.10	16.6 (2.0, 31.3)	**0.03**	8.13 (-7.3, 23.5)	0.26
Pause 3	15.6 (-5.9, 37.2)	0.10	-12.5 (-34.1, 9.1)	0.33	3.1 (-18.5, 24.7)	0.93
SCENARIO B	**Dr and SN**		**SN and JN**		**Dr and JN**	
	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value	Mean difference % (95% CI)	p value
Pause 1	17.3 (-2.4, 36.9)	0.09	2.7 (-16.9, 22.4)	0.94	20.0 (0.33, 39.7)	**0.05**
Pause 2	10.3 (-10.8, 31.4)	0.46	6.8 (-14.2, 27.9)	0.71	17.9 (-3.9, 38.2)	0.13
Pause 3	-13.6 (-34.6, 7.3)	0.26	11.4 (-9.6, 32.3)	0.34	-2.3 (-23.2, 18.7)	0.96

Table 8-7: One-way ANOVA for comparison of mean SAGAT scores between professional groups (*post-hoc test used Games-Howell correction)), significant results are highlighted in red. Dr = doctor, SN = senior nurse, JN = junior nurse.

SAGAT accuracy scores between team members in scenario A showed differences at pause 1 where junior nurses had significantly higher scores and pause 2 where senior nurses had significantly higher scores than junior nurses.

In scenario B the only significant difference occurred at point 1 between where the doctors had higher scores than the junior nurses. These are combined SAGAT scores (i.e. do not differentiate between levels of SA) Figure 8-2 revealed that the greatest difference between the doctors and the nurses in the more complex scenario B was in comprehension (level 2 SA) –

again this may be due to lack of communication of the doctors' understanding of the situation to the nurses.

8.5.2.6 ANALYSIS OF DIFFERENCE IN SART SCORES BETWEEN SCENARIOS

SART questionnaires were completed for each of the scenarios by all candidates except for team 9 where the SART forms for scenario A were inadvertently excluded from the debrief session, and six data points (four from scenario A and two from scenario B) which were not recorded by the participants.

This analysis was undertaken after the Cronbach's alpha calculation (see below) which revealed poor internal consistency for all domains in SART. Results are shown for the SART domains only. The calculation for SART was not done because of the reasons described in the above section on content validity and because of the poor internal consistency . Results are shown in Figure 8-4.

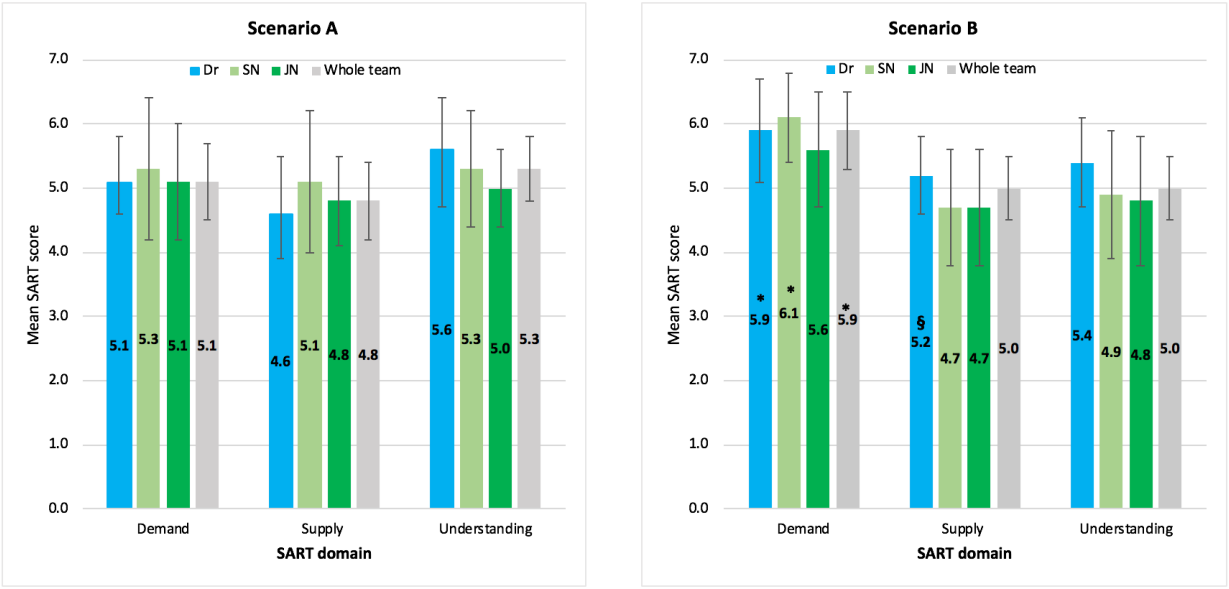


Figure 8-4: Mean SART domain scores for Scenario A and B. Scores were significantly higher in the demand domain for doctors, senior nurses and the whole team (* p < 0.03) in scenario B and in the supply domain for doctors (§ p = 0.04)

Data were normally distributed for both scenarios as assessed by Shapiro-Wilk's test (p = >0.05).

Mean SART scores were significantly higher in the demand domain in scenario B for doctors (mean difference 0.8, 95% CI 0.3 to 1.3, p=0.002), senior nurses (mean difference 0.75 95% CI 0.1 to 1.4, p=0.02) and the whole team (mean difference 0.72, 95% CI 0.3 to 1.1, p <0.0005). Scores were also significantly higher in the supply domain in scenario B for doctors (mean difference 0.5, 95% CI 0.04 to 1.0, p =0.04). There were no significant differences in mean SART scores for junior nurses or teams between scenario A and scenario B. These results provide very limited evidence for the ability of the SART tool to discriminate between an easier and a more complex scenario. The concerns raised when SART was assessed for content validity in a healthcare setting were reinforced by these data and by the data from the workload measure which will be described in Chapter 9 (i.e. that the "demand" domain in SART is more akin to workload than SA).

Analysis of internal consistency for SART was also poor (see below). Whilst poor internal consistency can explain the lack of discriminant validity it cannot define the cause: it may be that SART is not measuring SA in this context or that participants are using the tool differently. Further research will be necessary to define the role (if any) for SART in AICU settings.

8.5.2.7 ANALYSIS OF DIFFERENCE IN ANTS SCORES BETWEEN SCENARIOS

Total SA scores were used for analysis. Results for ANTS were normally distributed for all category scores and global scores in both scenarios (Shapiro-Wilk > 0.05) except for teamwork in scenario B. Results are shown in Table 8-8.

ANTS scores: categories and elements (score: 1-4)	Scenario A Mean (SD)	Scenario B Mean (SD)
Task management overall	3.2 (0.8)	3.1 (0.7)
Planning and preparing	3.3 (0.6)	3.4 (0.5)
Prioritising	3.2 (0.8)	3.2 (0.8)
Maintaining standards	3.0 (0.7)	3.2 (0.6)
Utilising resources	3.2 (0.7)	3.0 (0.7)
Teamwork	3.2 (0.6)	3.4 (0.5)
Coordination	3.3 (0.7)	3.2 (0.5)
Exchanging information	3.3 (0.7)	3.3 (0.7)
Authority and assertiveness	3.3 (0.7)	3.2 (0.6)
Assessing capabilities	2.8 (0.7)	3.2 (0.5)
Supporting others	3.3 (0.7)	3.4 (0.6)
Situation Awareness	3.2 (0.6)	3.2 (0.8)
Level 1	3.4 (0.6)	3.2 (0.8)
Level 2	3.3 (0.7)	3.2 (0.8)
Level 3	3.3 (0.7)	2.9 (0.9)
Decision making	3.3 (0.5)	3.0 (0.8)
Identifying options	3.3 (0.6)	3.1 (0.8)
Balancing risks	3.0 (0.6)	3.0 (0.7)
Re-evaluating	3.4 (0.5)	3.1 (0.8)

Table 8-8: Mean ANTS scores from one rater (HH) for categories and elements in scenario A and B in the SASi study

Independent measures t-tests revealed no significant difference in mean ANTS scores between scenario A and B for any category or element. The ANTS scoring system goes from 0-4 and problems with ceiling effect have been noted in this thesis (Chapter 6) and by Jepsen et al[312] when they adapted ANTS to a Danish setting. It may be that a larger scale would allow greater sensitivity in this context.

8.5.3 ANALYSIS OF RELATIONSHIP BETWEEN SA MEASURES

Pearson's product-moment correlations were used for normally distributed data sets and Spearman's correlation for non-normally distributed data to analyse the strength and direction of the relationship between SAGAT and ANTS for the whole team, the senior team and the professional groups. SAGAT overall scores, scores at SA levels and scores at pause points were

compared with ANTS global score (which incorporates SA) and the SA category scores for both scenarios.

SART was excluded from this analysis as internal consistency testing (see below) revealed that it was unlikely to be measuring SA.

No significant correlations were found between SAGAT and ANTS for any professional group or team combination. This is the first time SAGAT has been used in healthcare in conjunction with other measures of SA, it is not possible, therefore, to compare the SASi results with others. It is probable that the difference in the technique of measuring SA is implicated here, i.e. SAGAT is a direct measure and ANTS is indirect; SAGAT is scored on an individual basis whereas ANTS assesses the anaesthetist's SA in the context of working within a team and sharing information. The results also highlight the challenges in using observer-based tools to measure cognitive skills.[428]

8.5.4 RELIABILITY OF SA MEASURES

8.5.4.1 ANALYSIS OF AGREEMENT OF SAGAT SCORES WITHIN PROFESSIONAL GROUPS

Fleiss's kappa was used to determine if there was agreement in SAGAT answers within professional groups in scenarios A and B. The results are shown in Table 8-9.

Professional group	Agreement (κ) - Scenario A	Agreement (κ) - Scenario B
Doctors	0.44 (p<0.005) = moderate	0.16 (p<0.005) = poor
Senior nurses	0.35 (p<0.005) = fair	0.25 (p<0.005) = fair
Junior nurses	0.15 (p<0.005) = poor	0.26 (p<0.005) = fair

Table 8-9: Fleiss's kappa results for SAGAT answers by each professional group in SASi

Agreement was poor to moderate across professions in both domains.[387] SAGAT questionnaires are completed in the moment by individuals. The analysis was undertaken by professional group to determine levels of agreement in groups of healthcare professionals with similar levels of experience and the same clinical background. It is possible that the variability in experience of the doctors and cultural differences amongst participants (see Chapter 9) impacted levels of agreement. This will be discussed further below.

8.5.4.2 SART INTERNAL CONSISTENCY

SART has not previously been used to measure SA in a healthcare setting. Internal consistency is a measure of reliability and is an assessment of how much the items on a scale are measuring the same underlying dimension. Cronbach's alpha is the most commonly used measure of internal consistency and it was calculated for both scenarios. The results are shown in Table 8-10.

SART domain	Scenario A		Scenario B	
	Squared multiple correlation	Cronbach's alpha if item deleted	Squared multiple correlation	Cronbach's alpha if item deleted
Demand – Instability of situation	0.040	0.780	0.112	0.790
Demand – Variability of situation	0.416	0.323	0.497	0.026
Demand - Complexity of situation	0.433	0.152	0.461	0.411
Cronbach alpha - Demand	**0.580**		**0.549**	
Supply - Arousal	0.037	0.458	0.231	0.176
Supply – Spare mental capacity	0.040	0.389	0.074	0.574
Supply - Concentration	0.180	0.222	0.217	0.156
Supply – Division of attention	0.181	0.269	0.140	0.397
Cronbach alpha - Supply	**0.414**		**0.401**	
Understanding – Information quantity	0.300	0.383	0.151	0.478
Understanding – Information quality	0.202	0.585	0.210	0.281
Understanding - Familiarity	0.175	0.617	0.104	0.536
Cronbach alpha - Understanding	**0.631**		**0.538**	

Table 8-10: Results of tests of internal consistency (Cronbach alpha scores) for SART domains in scenario A and B. None of the scores reaches an acceptable level (i.e. all are <0.7)

The Cronbach alpha results reveal unsatisfactory internal consistency (i.e. <0.7[385]), across all domains for SART in this study. These results reveal that the SART tool used in this context is unlikely to be measuring SA. The "understanding" domain had the highest Cronbach alpha when both scenarios were combined and it is likely that this domain is most relevant to SA[429]. The authors of SART recognise that further adaptations to the tool may be necessary to clarify the domains and provide an overall score that represents SA alone[430].

8.5.4.3 ANTS INTERNAL CONSISTENCY

Internal consistency was analysed using Cronbach's alpha for ANTS in both scenarios, the results are shown in Table 8-11.

Category	α - Scenario A	α - Scenario B	α - Scenarios A+B
Task Management	0.81	0.84	0.82
Teamworking	0.87	0.89	0.88
Situation Awareness	0.75	0.93	0.88
Decision Making	0.62	0.91	0.84

Table 8-11: Results of internal consistency assessment (Cronbach's alpha) for ANTS scores in each category calculated for scenarios A and B separately and for both scenarios combined

Results for every category when both scenarios were combined were good (Cronbach's alpha >0.7). Where scenarios were considered separately the decision-making domain in scenario A was the only one which did not possess satisfactory internal consistency. These scores were generated by the author who has extensive experience in using ANTS and so it is reassuring that internal consistency was high with this tool.

8.5.5 USABILITY OF SA MEASURES

8.5.5.1 SAGAT

The GDTA required to generate SAGAT questions for two scenarios required a team of 10 subject matter experts and took two months to complete (including the answer template).

During the testing of the scenarios it was apparent that the use of our SAGAT questionnaire would add a significant amount of time to the training session and it was, therefore, only possible to apply SAGAT questions to one scenario which limited the amount of data we could collect.

Time taken to answer questions during the SAGAT pauses was measured in seconds for each team in both scenarios but is shown Table 8-12 in minutes and seconds for clarity.

Pause point	Scenario A mean pause length: min:sec (SD)	Scenario B mean pause length: min:sec (SD)
Pause 1*	08:09 (01:29)	09:53 (01:24)
Pause 2	05:51 (00:58)	06:45 (01:18)
Pause 3	03:58 (01:09)	04:16 (00:50)

Table 8-12: Pause times for SAGAT scenarios A and B. * Pause 1 was significantly longer in scenario B than scenario A

Independent samples t-tests were used to determine if there were significant differences in pause times. Homogeneity of variance (as assessed by Levene's test) was not violated for any data set.

Average pause length overall for both scenarios combined was 393 seconds (±77). Pause 1 was significantly longer in SAGAT Scenario B, a difference of 104 seconds (95% CI 20.1 to 189.0), $t(17) = 2.6$, $p = 0.02$. There was no statistically significant difference in pause times 2 or 3 for scenarios A and B.

8.5.5.2 SART

No adaptations were required from the original SART questions. A brief, standardised description of how to answer the questions was provided at the start of each training session and all candidates answered them without difficulty after each scenario. There were six data points missing (four from scenario A and two from scenario B) from the questionnaires (participants had not scored them) and three questionnaires were inadvertently not given to one team after they had completed scenario A. Therefore, a total of 1,104 data points were collected from a possible 1,140. This assessment of usability is included for completeness. Even though it is easy to use, SART would not appear to be a useful measure of SA in this context.

8.5.5.3 ANTS

The usability of ANTS has been described in Chapter 6.

8.6 DISCUSSION

This study has investigated the validity of three SA measurement tools in individuals and teams in simulated emergency scenarios and is the largest study to date of the measurement of SA in real clinical teams on AICU.

8.6.1 EVIDENCE OF VALIDITY FOR SA MEASURES

8.6.1.1 CONTENT VALIDITY

8.6.1.1.1 SAGAT

The iterative process of development of the GDTAs, questionnaires and answer templates required a substantial time commitment for 10 subject matter experts. When the process was complete, however, we had two GDTAs for recognised AICU emergencies which could be used repeatedly for many teams from AICU in the OUHT but could also be shared more widely across the NHS. Similar processes for additional scenarios can now be defined more clearly including an understanding of the work involved. Other groups have highlighted the challenges in designing SAGAT questionnaires and have opted to reduce the number of questions asked[136,150,409] or confine them to level 1 SA[149] to simplify the process. This may be acceptable for formative learning situations but inadequate data collection would prevent statistical analysis of the validity of SAGAT or its value as a tool to inform design of training interventions.

8.6.1.1.2 SART

The SART questionnaire was deemed appropriate in its original format for use in these scenarios with the proviso that the research team felt that the demand and supply domains were more aligned with workload than SA (see above). Subsequent analysis of internal consistency reinforced those concerns as explained above. SART was developed by context specific experts in military aviation and although it was designed to be applied in other work environments our results suggest that it does not have the same relevance in healthcare.

8.6.1.1.3 ANTS

Face and content validity of ANTS has been [152,267,312] studied by other research groups and by the author in this thesis.

8.6.2 EVIDENCE OF DISCRIMINANT VALIDITY

8.6.2.1 SAGAT

The ability of the SAGAT tool to discriminate was tested between scenarios, professional groups, SA levels and pause points.

Overall mean SAGAT accuracy scores for both scenarios were greater than 60% (scenario A = 70.1% ±7.3 and scenario B = 62.3% ± 14.4) and these scores were significantly higher in the easier scenario (A). However the difference was only 7.8% ±8.3 and the resolution of the scale (22 points) equated to approximately 4% per question which means that a statistically significant difference was found where they may be only 2 correct answers' difference between all participants.

Established teams work differently when compared with ad-hoc teams and this can lead to better or worse performance (e.g. established teams may perform better because they know their environment but ad hoc teams are not subject to the same interpersonal conflicts that may exist amongst staff who work together every day).[116,431] Our teams were taken directly from AICU and their familiarity with each other and with procedures on AICU may have contributed to better SA in both scenarios. Measures of performance will be explored in more depth in the following chapter.

Analysis of SAGAT accuracy at different levels of SA revealed a significant increase between level 1 SA and Level 3 SA for all professions in both scenarios. This may not make intuitive sense at first glance because gathering of adequate information at level 1 should be a pre-requisite for the development of comprehension at level 2 and the ability to project at level 3. The explanation may lie in the way we see data and the fact that humans will gain the "gist" of a scene or unfolding scenario without actually memorising accurate details.[432] It is possible that asking the question "what was the blood pressure" may not be the appropriate way to assess level 1 SA in these scenarios but that something like "was the mean arterial pressure above 65 mmHg ?" would be better.

The only significant difference observed between professions at levels of SA occurred at level 2 (comprehension) in the more complex scenario (B) where the doctors had a mean SAGAT accuracy score 20% higher than both the senior and the junior nurse. There were eight requirements for comprehension defined in the GDTA for scenario A and 11 in scenario B and both respiratory and cardiovascular signs were important in scenario B whereas problems were limited to the cardiovascular system in scenario A. This provides some evidence to support discriminant validity at the level of comprehension between a simpler and a more complex

scenario although Endsley[147] does caution that combining queries into groups at e.g. SA level may reduce the sensitivity of the metric due to shifts in attention and variability in demands of the task in hand. However, SAGAT questions were asked in standardised scenarios at time points dictated by the same parameters for every team in order to reduce the impact of such variations.

Discriminant validity of SAGAT at pause points was confirmed by a predictable rise in SAGAT accuracy with time as participants had more time to gather and share information. The junior nurse entered the room first in scenario A followed by either the doctor and the senior nurse together (this happened in five of eight scenarios) or the senior nurse and then the doctor, when the junior nurse asked for assistance. The analysis showed significantly higher mean SAGAT accuracy scores for the junior nurses versus the doctors at pause point one in scenario A providing evidence that SAGAT could discriminate between participants who had more time to develop SA. The reverse was true in the more complex scenario (B) where all staff began the scenario together and the doctors had significantly higher mean SAGAT scores than the junior nurses at pause 1. It is possible that mental schemata (which provide templates for comparison of situations[61]) for the events in the more complex scenario B were better developed in doctors than junior nurses.

8.6.2.2 SART

SART did not discriminate between scenarios or participants. The analysis of internal consistency revealed low Cronbach's alpha scores suggesting that SART is measuring a different construct from SA which explains this finding. Lack of evidence of validity for SART may also be impacted by the effect of personal bias which is evident in many self-assessment tools.[151]

8.6.2.3 ANTS

Both the global ANTS and the SA domain scores were used to test for differences between the two scenarios and neither was able to detect a difference. ANTS is an observer-based tool designed to measure NTS in anaesthetists in any clinical setting. ANTS was chosen as the observer based tool in this study because overall IRR was very good and it was designed to be used in a variety of clinical contexts. However, the study of NTS tools in chapter 5 revealed that OSCAR[336] had the highest IRR scores for the SA domain. This may be because OSCAR focuses on one specific problem – cardiac arrest. It is possible that using OSCAR as a measure of team SA at the point of arrest in both scenarios would provide greater discrimination but this would be at the expense of assessing SA across a whole scenario and it could not be used for scenarios where the management cardiac arrest was not assessed.

8.6.3 RELATIONSHIPS BETWEEN THE SA MEASURES IN SASI

Evidence of validity in terms of relationship with other measures of the same basic construct (sometimes called concurrent validity)[139] revealed no correlations between SAGAT and either global or SA scores in ANTS.

Whilst these tools are measuring SA they are doing so in different ways: SAGAT is a direct objective measure taken from participants and ANTS relies on observations made by a trained educator based on observed behaviours. The results highlight the challenges faced in measuring SA using inferences from participant behaviours.

8.6.4 EVIDENCE OF RELIABILITY FOR SA MEASURES

8.6.4.1 SAGAT

Reliability was considered in terms of participant agreement on each of the 22 questions for each profession in both scenarios. Results from the SASi study are similar to those found by Morgan et al[150] in their study of three obstetric multidisciplinary teams managing emergency situations where the highest kappa score achieved was 0.57 (in this study the highest kappa was 0.44). This measure of agreement does not determine whether the answers were correct, merely that the answers were the same or different. The agreement in this study was mostly only fair to moderate. This could be methodological, if the questions were hard to answer or mistakes were made in interpreting the questions then agreement would fall. At a participant level this may be because there was inadequate sharing of information across teams, or that the SAGAT questions were not focused well enough on information relevant to SA at each level.

Agreement of SAGAT scores between team members i.e. shared SA,[433] has been highlighted as an important aspect of team performance. However, even more importantly an understanding of which answers are the same *and* correct this has been termed "true shared SA"[403] and may provide greater clarity on why errors in SA occur. Analysing the data for levels of shared SA for whole teams and team pairings may provide additional insights into how SA is shared but simple percent agreement would not account for agreement by chance and underlying SAGAT accuracy levels may confound the calculation i.e. high levels of accuracy would lead to higher levels of agreement.

8.6.4.2 SART

Items used to form a scale need to have internal consistency i.e. the items should measure the same thing and should be correlated with each other. Cronbach's alpha is the most commonly used statistical tool to assess internal consistency.[385,434]

Cronbach's alpha scores for all SART domains were <0.7 which is unsatisfactory for groups using a scoring system. Levels of >0.7 for research tools used to compare groups may be considered satisfactory but for clinical applications even higher values would be required (often >0.9[385]).

These results reveal that SART does not possess adequate internal consistency and may be measuring a construct unrelated to SA.

8.6.4.3 ANTS

Internal consistency scores for ANTS were good (> 0.8) for all categories when scenarios were combined and only the decision making category for scenario A scored < 0.65. This is important information about the tool itself in the context of the SASi study but if it is not able to discriminate between participants or scenarios the question is, is it as a useful measure of SA in this context?

8.6.5 USABILITY

8.6.5.1 SAGAT

The GDTA and generation of SAGAT questions for both scenarios involved ten subject matter experts in an iterative process. This took one month and required careful coordination with six revisions of GDTAs and SAGAT questions. Whilst the process required substantial effort, once

completed for a bank of scenarios it would not require too much input to keep the scenarios up to date. This may, however, be off-putting for small teams of simulation educators. It may require involvement of national organisations, such as ASPiH, to provide a library of such scenarios to allow greater use of SAGAT and more opportunity for research into SA in teams. A national library of scenarios would also allow comparative measures of team performance between simulation centres.

Concerns have been raised about the intrusiveness of the SAGAT pauses[84] Studies in military aviation[172] and nuclear power[435] found no evidence that the pauses impacted performance. More recently in healthcare, Morgan et al[150] considered the intrusiveness of the number of SAGAT questions asked by asking for feedback from participants who had experienced different numbers of queries at pauses in three scenarios. They found that 9 queries per stop was considered too many by 57% of participants, 6 queries was considered too many by 13% and no one found 3 queries intrusive. Pauses in the SASi study were longer than anticipated but feedback was similar: 86% (49/57) participants found SAGAT pauses not at all or mildly disruptive to training and only 14% (8/57) found them very disruptive.

8.6.5.2 SART

The SART questionnaire required no adaptation and was straightforward to administer and answer (as evidenced by low missing score rates). However, it does not appear to be a useful tool for the measurement of SA in healthcare settings.

8.6.5.3 ANTS

The usability of ANTS was assessed in Chapter 6.

8.7 STUDY LIMITATIONS

Constraints of service provision meant that teams were allocated to the training sessions on the morning of training. We could not randomise staff to attend on certain dates.

There was an unequal balance of the two SAGAT scenarios (8 teams used SAGAT questionnaires for scenario A and 11 for scenario B). Unfortunately, the unpredictable cancellation of sessions fell more often on days when SAGAT was randomised to scenario A.

It is recommended that at least 30 samplings are taken for each question[147] – this was achieved for scenario B (33 samplings per question) but not scenario A (24 samplings per question).

We did not have an equal balance of SAGAT questions at each of the pauses for the two scenarios – context of the scenario at the time prevented even distribution. We did, however, have a more even spread of questions at SA levels 1, 2 and 3 (8, 8 and 6 respectively)

Despite the fact that we had tested the time taken to fill in the SAGAT answer sheets and the NASA-TLX (which were completed at the same time) we did not make adequate adjustments for thinking time for the candidates. Results from the feedback questionnaire, however, revealed that the majority of participants did not feel that the SAGAT pauses disrupted their learning.

Designing questions which were unambiguous and easy to classify as directed at level 1, 2 or 3 SA was challenging – it is possible that this impacted the way they were answered by participants although a member of the research team was present with every participant whilst they answered questions in case there were any problems.

8.8 CONCLUSION AND FUTURE RESEARCH

This study has shown that the measurement of SA is challenging and that available tools are imperfect. Evidence of validity has been shown for SAGAT but not SART or ANTS in this context. SAGAT has been used in the design of avionics systems and air traffic control systems and has highlighted issues in the design of both.[147] There is limited evidence that it may be used in a similar way to inform training interventions.[168] We did not use SAGAT to investigate the human-device interactions relevant to the development of SA (such as data from the ventilator, infusion devices or the monitor). It is possible that using SAGAT to inform improvements in isolated aspects of the environment (such as human-monitor interfaces) may be more fruitful than using it to assess whole system performance because of the many competing demands on attention.

This study has lain the groundwork for objective measurement of SA in healthcare professionals working on AICU in the OUHT but more research is needed to understand the most informative measure of SA and the links between improved SA, teamwork, performance and outcome for patients.

The following chapter will present results from the SASI study which explored the relationship between experience, workload, team performance and SA.

CHAPTER 9 A STUDY OF SITUATION AWARENESS IN SIMULATION FOR ADULT INTENSIVE CARE (SASI) – THE RELATIONSHIP OF SA WITH EXPERIENCE, WORKLOAD AND PERFORMANCE

9.1 INTRODUCTION

Situation awareness can be measured reliably in healthcare settings and loss of SA has been a critical determinant of error in high profile cases. Workload, experience and performance have been found to correlate with SA in military and other work settings. More experienced personnel have increased SA when compared with less experienced[436], lower cognitive workload leads to increased SA[437] and errors in SA have been negatively correlated with performance.[109] More recently, in healthcare, these correlations have been explored in simulated settings in trauma,[136,410] emergency medicine,[149] ICU[420] and obstetrics[150] and with medical students[418,421] and nursing students.[409,417]

The SASi study described in the previous chapter is the first to use real teams of ICU staff with standardised simulated scenarios. Studies in healthcare have assessed anaesthetists and bioengineering students in simulated scenarios to analyse monitor design;[148] doctors in simulated scenarios;[420] doctors and medical students in simulated scenarios;[136] medical students only in simulated scenarios;[418,421] nursing students only in simulated scenarios;[409,438] and doctors in a real ED setting.[149] Only two studies have looked at the use of SAGAT in MDTs,[150,410] both were smaller studies (four and three teams respectively) and neither was in ICU.

The SASi study was designed primarily to consider the validity of SAGAT, SART and ANTS as measures of SA in simulated scenarios for AICU teams. Evidence of validity (content and relations with other measures [discriminant and concurrent validity]) of these tools has been described in the preceding chapter. This chapter will now consider relations with similar, related constructs i.e. the correlation of SA measures with: experience, technical performance and workload which has important implications both for clinical practice and training.

Objectives and outcome measures for this study are described in Table 9-1.

	Secondary Objectives	Analysis of Outcome Measures
1	To investigate the effect of staff experience and professional background on maintenance of SA	Correlation of experience and professional background with measures of SA
2	To study the effect of workload on SA in AICU simulation training	Correlation of NASA TLX data with measures of SA
3	To investigate the relationship between SA and technical performance in teams of AICU staff	Correlation of SA measures with technical performance (as measured by time and global performance scores)

Table 9-1: SASI study: Secondary objectives and their outcome measures

9.2 METHODS

The design, participants and procedures have already been described in the previous chapter. The methods section in this chapter will present the measures used to assess experience, workload and performance which were not previously described.

9.2.1 MEASURES

9.2.1.1 PARTICIPANT EXPERIENCE LEVEL

Experience was measured as years of working in AICU. All doctors were in training and ICU experience was counted as specific blocks of specialist training in ICU added together. Where

participants had less than one year's experience this was recorded as experience in months and calculated as a percentage of one year.

All doctors were in training grades. There were two nurses in every team and, wherever possible, one was more experienced than the other such that each team had a senior and a junior nurse.

9.2.1.2 NASA TLX

Assessment of workload was undertaken during the SAGAT scenarios using the NASA-TLX score[425] and was recorded by each participant on an iPad (using the app: https://humansystems.arc.nasa.gov/groups/TLX/tlxapp.php) at the three pause points in each SAGAT scenario. Participants provided scores (on a scale of 0-100) relating to workload across the six domains in the TLX system (mental demand, physical demand, temporal demand, performance, effort and frustration) as well as an assessment of the importance of each domain in the context of the preceding action which allowed calculation of a "weighted" score for each domain and overall. The NASA-TLX score sheet and an aide memoire describing each domain (which was provided on paper for each participant to refer to if necessary) are included in Appendices 29 and 30.

In a review of the use of the NASA-TLX[439] the original author found that the most common modification made to the score was to eliminate the weighting altogether and measure only raw data and states that studies using the "Raw TLX" and comparing it to the original found it "…either more sensitive,[440] less sensitive[441], or equally sensitive,[442] so it seems you can take your pick."

In order to decide whether to use raw or weighted data I first conducted an analysis of the internal consistency of the NASA-TLX scores in raw or weighted states and the results are described below.

9.2.1.3 TEAM PERFORMANCE MEASURES

9.2.1.3.1 GLOBAL TECHNICAL PERFORMANCE SCORE (GPS)

Global technical performance scores have been shown to possess greater construct validity and better reliability than checklists. Furthermore, when administered by experts, they are a more appropriate measure of performance.[443,444] A GPS was designed to assess global team performance of tasks using a measurement scale of 0-100 (i.e. the same as the NASA TLX scale). The development of the GPS was undertaken by the same team of experts involved in the SASi study and was informed by best practice guidelines in performance measurement in simulation training.[370] The scale chosen was the same as the NASA-TLX (to reduce the risk of ceiling effect and to provide continuity) and this scale was applied to team performance by HH and LV based on the expected actions defined in the GDTAs for both scenarios. The score sheet with instructions on how to apply it to both scenarios is provided in Appendix 31.

Post-hoc video analysis has been found to be a reliable means of assessing technical and non-technical performance in healthcare[445–447] and confers some advantages including reduced observer distraction and the ability to replay key sections of the scenario.

Scores were, therefore, generated after video review by HH and LV and intraclass correlation was calculated to assess inter-rater reliability.

9.2.1.3.2 SCENARIO TIMINGS

Time taken to complete tasks has been used a measure of performance in simulated environments in studies of intubation,[448] obstetric crisis management[150] and laparoscopic surgery.[449] Key time points were measured for scenarios A and B to provide an additional quantitative measure of performance. Times taken to complete the whole scenario (from participants entering the simulation room to the point that faculty stopped the action) as well as times for each phase of the scenario were recorded. A "total action period" (TAP) for each scenario was defined as the point at which patient deterioration began to the point the final treatment was given. This meant that the time of onset and completion of each team's interaction with the patient was standardised for all teams and the initial and ending few moments in the scenario (where there was not much activity or resolution of treatment had already occurred) were removed.

Analysis was undertaken initially for all action periods (i.e. the three phases of each scenario, as described in the previous chapter, and overall) but the initial results suggested that this was not providing any additional useful information and I have, therefore, presented only the results comparing SA with TAPs.

9.2.2 USABILITY

We assessed usability of the TLX system by recording times taken in the pauses to complete the SAGAT and NASA-TLX questionnaires combined.

9.2.3 MEASUREMENTS DURING THE STUDY

Table 9-2 is provided from the previous chapter as a reminder.

Timing of measurement	Measurement tool
Measurements before scenarios	Feedback questionnaire: pre-training questions for 57 participants **including participant experience level**
Measurements during scenarios	SAGAT questionnaire for 19 scenarios (used in each session for one scenario) **NASA-TLX for 19 scenarios** Observer based assessments of NTS including SA using ANTS *to inform the debrief* for all scenarios (not formally scored)
Measurements after scenarios	SART questionnaire for 57 participants (both scenarios) Feedback questionnaires: post-training questions and free text for 57 participants
Measurements after training completion	**Observer-based assessment of global technical performance for teams in 38 scenarios (HH and LV)** Observer-based scoring of NTS including SA using ANTS for 38 scenarios (HH) **Measurement of total action period for 38 scenarios (HH)** Scoring of SAGAT questionnaires (HH)

Table 9-2: Timing of all measurements taken during training sessions and afterwards for the SASi study. Measures discussed in this chapter are in bold script and underlined.

9.2.4 DATA ANALYSIS

Assessment of normality of distribution was made for the NASA TLX data and for performance measures (both GPS and TAPs) and mean values were calculated for comparisons. Where data were not normally distributed the appropriate test for non-parametric data was used instead. Outliers were detected by assessment of boxplots, their presence is noted, where relevant, in the results and all were retained in analyses.

9.2.4.1 ANALYSIS OF RELIABILITY FOR NASA-TLX AND GPS

Internal consistency of the NASA-TLX was measured using Cronbach's alpha. Interrater reliability (IRR) of the GPS (for HH and LV) was measured with intraclass correlation. Internal consistency and IRR for ANTS have been described previously.

9.2.4.2 ANALYSIS OF VALIDITY (DISCRIMINANT)

Independent measures t-tests were used to analyse differences in workload and for performance measures (both GPS and TAPs) between scenarios (Welch's t-test was used if data were not normally distributed), professional groups and teams. Repeated measures ANOVA was used to measure differences in workload between professional groups in both scenarios.

9.2.4.3 ANALYSIS OF RELATIONS WITH OTHER VARIABLES

Measures of SA (SAGAT and ANTS) were compared with measurements of related constructs namely experience of participants, workload (NASA-TLX) and technical performance (GPS and TAPs).

Pearson product-moment correlation coefficients were used for normally distributed variables and Spearman's rank-order correlation for non-normally distributed variables. Independent measures t-tests to analyse difference in assessment scores between scenarios and professional groups and RM ANOVA for measurement of differences between more than two variables.

9.3 RESULTS

9.3.1 EXPERIENCE LEVEL OF PARTICIPANTS

Average experience on AICU (in years) for 19 doctors was 3.6 (± 2.3); 19 senior nurses 4.2 (± 1.7) and 19 junior nurses 1.7 (± 1.1). Average experience in years was calculated for the whole team: 9.5 years (± 3.1) and the senior team 7.8 years (± 2.6).

Experience levels of the teams taking part in the SAGAT scenarios were greater for Scenario B than scenario A as shown in Figure 9-1 and Table 9-3.

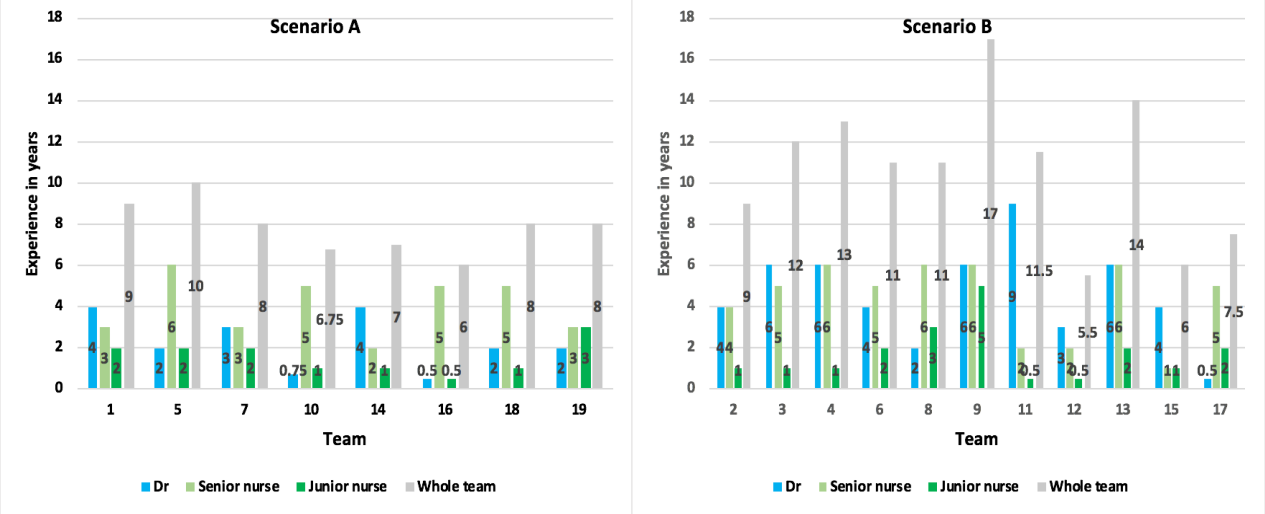


Figure 9-1: Experience levels (in years or fractions of years) for professional groups and whole teams in SAGAT scenarios A and B

The doctors in eight of the teams (5, 6, 8, 10, 16, 17, 18 and 19) were less experienced than the senior nurses. The least experienced team (12) had 5.5 years of experience and the most experienced (9) had 17 years of experience working in AICU.

Staff group	SAGAT Scenario A Average experience years (SD)	SAGAT Scenario B Average experience years (SD)
Whole team	7.8 (1.3)	10.7 (3.5)
Senior team	6.3 (1.0)	8.7 (2.9)
Doctors	2.3 (1.3)	4.6 (2.1)
Senior nurses	4.0 (1.4)	4.4 (1.9)
Junior nurses	1.6 (0.8)	1.7 (1.3)

Table 9-3: Mean experience levels (in years) for professional groups and teams in SAGAT scenario A and SAGAT scenario B

Independent measures t-tests revealed higher levels of experience in SAGAT scenario B for doctors (but not senior or junior nurses) a mean difference of 2.3 years (95% CI 4.2 to 0.4), t(17) = 2.512, p=0.02. When combined experience for whole teams was analysed (using Welch's t-test) higher levels of experience were also found for Scenario B with a statistically significant

mean difference of 2.8 years (95% CI 5.3 to 0.4) t(13.389) = 2.489, p=0.03) and for the senior team with a mean difference of 2.7 years (95% CI 4.7 to 0.7) t(13.007) = 2.887, p=0.01).

The greater levels of experience of the doctors in scenario B influenced the team experience levels and it is possible that this led to a less obvious difference between the simpler (A) and more complex (B) scenario.

9.3.2 WORKLOAD MEASURE: NASA-TLX

9.3.2.1 RELIABILITY OF NASA-TLX

Reliability (internal consistency) was measured for all participants and for professional groups in both scenarios combined using raw and weighted data. The majority of Cronbach's alpha results were negative when weighted scores were analysed i.e. there was a negative average covariance which violates reliability model assumptions.[450] Further analysis of weighted scores was, therefore, abandoned and only raw scores have been considered.

Results are shown in Table 9-4. Items coloured black (i.e. mental demand, temporal demand, effort and frustration) would have *decreased* the overall internal consistency of the tool if removed from the analysis. Pearson's correlation coefficients are presented in parentheses for items coloured red (removal of these items would *increase* the internal consistency of the tool). If the value is less than 0.3 the item may not be measuring the same underlying construct (i.e. workload in the case of the NASA-TLX) and should probably be removed from the analysis.

NASA-TLX domain	All participants	Doctors	Senior nurses	Junior nurses
Cronbach's alpha	**0.69 (0.73)**	**0.75 (0.81)**	**0.71 (0.73)**	**0.61 (0.68)**
	Cronbach's alpha if item deleted			
Mental demand	0.64	0.68	0.67	0.55
Physical demand	0.70 (0.37)	0.77 (0.42)	0.70	0.64 (0.26)
Temporal demand	0.62	0.68	0.67	0.51
Performance	0.72 (0.16)	0.78 (0.17)	0.73 (0.26)	0.66 (0.07)
Effort	0.61	0.67	0.61	0.53
Frustration	0.62	0.70	0.66	0.49

Table 9-4: Results of internal consistency tests (Cronbach's alpha) for NASA-TLX raw scores. Values in parenthesis on the top row are recalculated Cronbach's alpha scores after items in red have been removed.

Cronbach's alpha scores of greater than 0.7 were achieved in all groups (except junior nurses which was close at 0.68) after successive removal of physical demand and performance categories (as the analysis revealed that their removal would improve the overall Cronbach alpha for the scale). Subsequent analysis of NASA-TLX scores was, therefore, restricted to mental demand, temporal demand, effort and frustration and a TLX "overall score" was calculated as the average of those four domains for each participants and the 19 teams. Results are shown in Figure 9-2.

9.3.2.2 EVIDENCE OF VALIDITY (DISCRIMINANT)

Independent measures t-tests were used to analyse differences in overall NASA-TLX scores (for all domains combined ["overall score"] and for individual domains considered separately) between scenarios when scores were averaged across the three pause points. Tests of normality revealed data were normally distributed for all participants in both scenarios.

9.3.2.2.1 EVIDENCE OF VALIDITY FOR NASA-TLX SCORES AVERAGED ACROSS THE WHOLE SCENARIO

There was no significant difference in overall NASA-TLX scores between scenario A and scenario B when all participants were considered together. However, significantly higher overall TLX (15

point increase, p=0.01), average temporal demand (13 point increase p=0.04) and average frustration scores (18 point increase p=0.04) were found for doctors in scenario B (the more complex scenario) compared with scenario A. Furthermore, scores were significantly higher for average frustration in scenario B compared with scenario A when whole teams were analysed together. This provides evidence that the NASA-TLX can discriminate between higher workloads in these two scenarios.

9.3.2.2.2 EVIDENCE OF VALIDITY FOR NASA -TLX SCORES AT EACH PAUSE POINT

One-way repeated measures ANOVA tests were conducted to determine whether there were statistically significant differences in overall NASA-TLX scores over the course of the scenarios (i.e. when measured at each pause) within professions. Data for Scenario A were normally distributed (as assessed by Shapiro-Wilk p>0.05) with 5 outliers in total. Data were also normally distributed for scenario B (as assessed by Shapiro-Wilk p>0.05) with 15 outliers in total. All outliers were included in the analyses. A Bonferroni post hoc test for multiple pairwise comparisons (i.e. between pause points) was applied and results are shown in Figure 9-2 and Table 9-5.

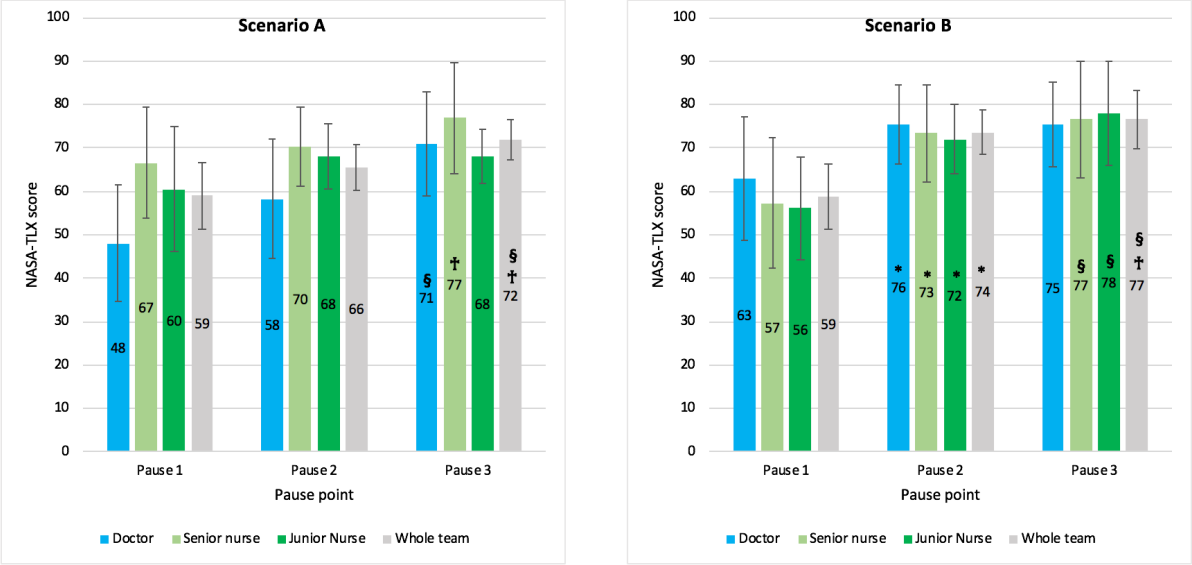


Figure 9-2: Overall TLX scores at pause points in scenarios A and B. Significant increases in TLX scores are highlighted at pause 1 versus pause 2 (* p ≤ 0.02), at pause 2 versus pause 3 († p = 0.03) and pause 1 versus pause 3 (§ p ≤ 0.02)

SCENARIO A	P1-P2		P2-P3		P1-P3	
	Mean difference (95% CI)	p value	Mean difference (95% CI)	p value	Mean difference (95% CI)	p value
Doctor	-9.5 (-27.4, 8.3)	0.4	-12.7 (-30.3, 5.0)	0.2	-22.2 (-37.5, -6.9)	**0.008**
Senior nurse	-3.6 (-14.8, 7.6)	1.0	-6.7 (-12.9, -0.5)	**0.03**	-10.3 (-24.7, 4.10	0.2
Junior nurse	-7.5 (-18.8, 3.9)	0.2	0 (-5.9, 5.9)	1.0	-7.5 (-19.5, 4.5)	0.3
Whole team	-6.7 (-14.4, 1.1)	0.09	-6.3 (-11.8, -0.8)	**0.03**	-13.0 (-23.7, -2.3)	**0.02**
SCENARIO B	**P1-P2**		**P2-P3**		**P1-P3**	
	Mean difference (95% CI)	p value	Mean difference (95% CI)	p value	Mean difference (95% CI)	p value
Doctor	-12.6 (-23.0, -2.2)	**0.02**	0.2 (-5.8, 6.3)	1.0	-12.4 (-25, 0.2)	0.06
Senior nurse	-16.1 (-25.3, -7.0)	**0.001**	-3.2 (-10.1, 3.8)	0.7	-19.3 (-26.1, -12.5)	**0.001**
Junior nurse	-15.8 (-27.2, -4.3)	**0.008**	-6.0 (-14.8, 2.7)	0.2	-21.8 (-37.9, -5.8)	**0.009**
Whole team	-14.8 (-18.4,-11.3)	**<0.005**	-3.0 (-5.6, -0.1)	**0.04**	-17.8 (-22.7, -13.0)	**<0.005**

Table 9-5: Results of repeated measures ANOVA for analysis of differences in overall NASA-TLX scores within professional groups at different pause points in scenarios A and B in the SASi study

There were significant increases in NASA-TLX scores between the beginning and the end of scenario A for doctors, senior nurses and the whole team but not for the junior nurses. There were significant increases in NASA-TLX scores for all groups between the beginning and end of scenario B.

These results provide evidence for discriminant validity of the NASA-TLX tool in the SASi study because the scenarios became more difficult with time. Furthermore, increases in scores during the more complex scenario (B) were greater for all professions and whole teams than in scenario A and those differences achieved significance for all groups earlier in scenario B (i.e. by pause 2) than in scenario A.

9.3.3 MEASURES OF PERFORMANCE

9.3.3.1 GLOBAL PERFORMANCE SCORES (GPS)

Global performance scores were generated for the 38 scenarios using post hoc review of videos by HH and LV. Interrater reliability was tested and independent samples t-tests used to determine if there were differences between scenarios. Results are shown in Figure 9-3.

9.3.3.1.1 INTERRATER RELIABILITY FOR GPS

Interrater reliability was analysed for the technical performance scale in SASi and the ICC for the two raters (HH and LV) was 0.70 ($p<0.001$). This represented good agreement and, in order to reduce observer bias the performance scores from LV were used in the analysis (because the ANTS scores were generated by HH alone). Scores were not normally distributed in Scenario A (Shapiro-Wilk <0.05) but were normally distributed in scenario B (Shapiro-Wilk > 0.05). Median scores were used to compare the two scenarios (because of the difference in distribution of

data) and median team technical performance score for Scenario A was 70 (range 60-80) and for Scenario B it was 65 (range 50-85).

9.3.3.1.2 EVIDENCE OF VALIDITY (DISCRIMINANT) FOR GPS

Independent samples t-tests revealed no significant difference in mean technical performance scores overall between scenario A and scenario B or for SAGAT scenarios considered separately. This may be because overall team performance is good in the AICU at the OUHT and may also be impacted by the fact that in the harder scenario (B) the teams were led by doctors with more experience.

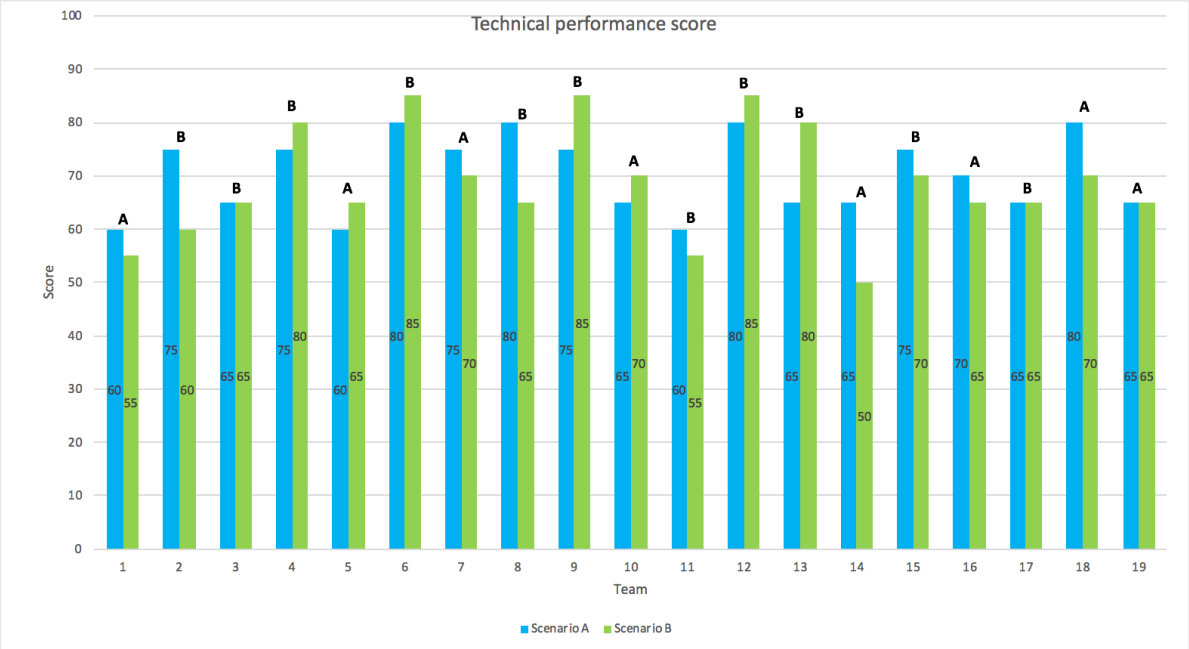


Figure 9-3:Technical performance scores for 19 teams in the SASi study (SAGAT scenarios are marked "A" or "B" for each team)

None of the teams scored less than 50 on the technical performance scale and only one (team 14) scored 50 for one scenario.

The overall high level of technical performance in these teams is reassuring but may explain, in part, the limited evidence of validity for the SA measures which were also good in both

scenarios for both teams (SAGAT scores >60% for whole teams in both scenarios and ANTS scores ≥ 3.0 for all categories in both scenarios).

9.3.3.2 TOTAL ACTION PERIODS (TAPS)

Overall scenario times were recorded automatically by the Simview software. Scenario A was longer than scenario B for 17 of the 19 teams. Results are shown in Table 9-6(for overall times and TAPs) and Figure 9-4 (for TAPs).

TAPs were calculated as described above for each team in both scenarios. All teams reached the prescribed end point in scenario A (resolution of asystolic cardiac arrest). However, in scenario B Team 1 did not reach the point of cardiac arrest (VF) because they were slower to make decisions earlier in the scenario and Team 5 did not pick up the pneumothorax or recognise ventricular fibrillation. A time penalty (10% added to the time taken by the slowest team) was given to TAPs in both cases.

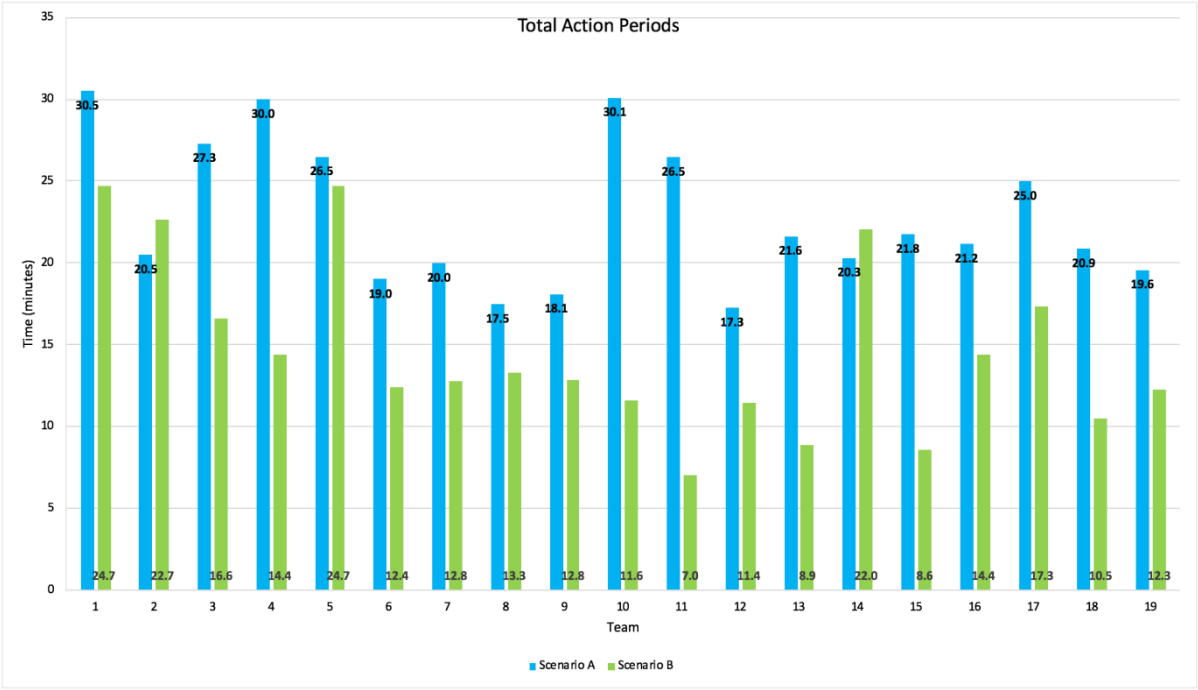


Figure 9-4: Total Action Periods (in minutes) for Scenarios A and B for each team in the SASi study

Three teams took similar lengths of time to complete both scenarios (teams 2, 5 and 14) the remainder took longer to complete scenario A. Statistical analysis revealed significant differences.

Time periods	Scenario A	Scenario B
Overall length of scenarios : secs (SD) Mins:secs (SD)	1653 (289)* 27:33 (04:49)	1361 (220) 22:42 (03:40)
Mean TAP for all scenarios : secs (SD) mins:secs [SD]	1360 (270)* 22:40 [04:30]	859 (270) 14:33 [04:30]
Mean TAP SAGAT scenarios: secs (SD) mins:secs [SD]	1403 (293)* 23:23 [04:53]	772(261) 12:52 [04:21]

Table 9-6: Mean Total Action Periods for scenario A and B with and without SAGAT questionnaires (significantly longer times are highlighted: *p≤0.001)

Data for times were normally distributed. An independent samples t-test was used to determine if there was a significant difference in TAPs between Scenario A and Scenario B. There was homogeneity of variances, as assessed by Levene's test for equality of variances: p=0.797. Mean TAPs were longer in scenario A (1360 seconds) than scenario B (859 seconds), as statistically significant difference of 500 seconds (95% CI 307 to 694), t(36) = 5.2, p < 0.0005

An independent samples t-test was used to determine if there was a significant difference in TAPs between Scenario A and Scenario B when SAGAT was used. There was homogeneity of variances, as assessed by Levene's test for equality of variances: p=0.384. Mean TAPs were longer in scenario A (1403 seconds) than scenario B (772 seconds), a statistically significant difference of 631 seconds (95% CI 362 to 900), t(17) = 4.9, p = <0.0005.

The easier scenario (A) took significantly longer than the more difficult one (B) which was surprising, but the explanation may be:

- The easier scenario (A) required focus on only one system (cardiovascular) and was reported by candidates in the debriefing sessions as feeling less urgent. This may have led the team leaders to take more time in their decision making and, therefore, lengthened the TAPs
- The more difficult scenario (B), had more decisions and tasks to undertake but heightened performance levels may be associated with additional pressure

9.3.4 RELATIONSHIP OF MEASURES OF SA WITH LEVEL OF EXPERIENCE, WORKLOAD AND PERFORMANCE

An assessment of evidence for construct (convergent) validity between SA measures and experience, performance or workload measures was confined to a comparison of SAGAT and ANTS (SART was excluded for reasons explained previously) with experience, the GPS, TAPs and the NASA-TLX (overall and for the domains of mental demand, temporal demand, effort and frustration).

Pearson's product moment and Spearman's correlation coefficients were used for parametric and non-parametric data respectively. Visual assessments of scatter plots revealed an absence of clear linear relationships (R^2 <0.5).

9.3.4.1 DIRECT MEASURE OF SA – SAGAT

Table 9-7 shows Pearson correlation coefficients depicting the relationship between SAGAT, experience, performance and workload for whole teams (data for the separate professions are included in Appendix 32.

Correlation with SAGAT score	Scenario A Pearson's *r* (p-value)	Scenario B Pearson's *r* (p-value)
Team SAGAT scores		
Experience	-0.404 (0.32)	-0.277 (0.41)
Workload (NASA TLX overall)	-0.190 (0.65)	0.271 (0.42)
Workload (NASA TLX MD)	-0.124 (0.77)	0.395 (0.23)
Workload (NASA TLX TD)	0.006 (0.99)	0.138 (0.69)
Workload (NASA TLX effort)	-0.321 (0.44)	0.285 (0.40)
Workload (NASA TLX frustration)	-0.073 (0.86)	0.184 (0.59)
Performance (GPS)	0.439 (0.28)	0.122 (0.72)
Performance (TAP)	0.141 (0.74)	0.464 (0.15)

Table 9-7: Results of analysis of correlations (Pearson's correlation coefficients) for whole teams between SAGAT score and NASA-TLX experience, performance and workload domains

These data reveal that no correlations were found between SAGAT scores for whole teams and the measures of participant experience, workload and performance (performance was measured as a whole team) in the SASi study. This may be attributed, in part, to the small number of teams analysed (19) but also to the fact that the tools used were individual scores (SAGAT and NASA-TLX) or whole team (GPS and TAPs). Further discussion will be provided below.

9.3.4.2 INDIRECT OBSERVER BASED MEASURE OF SA – ANTS

ANTS was designed for use as an observer-based measure of an anaesthetist's NTS and the doctors were scored in all 38 of the SASi scenarios. These observational scores were used as a surrogate indicator of team SA as the categories in ANTS are all dependent on team interactions. Correlations were found for SA measured using ANTS scores for doctors in scenario A (and in one category for scenario B where the ANTS global scores for the doctors were positively correlated with team performance) (see Table 9-8).

Correlation with ANTS score	Scenario A Pearson's *r* (p-value)	Scenario B Pearson's *r* (p-value)
ANTS SA score		
Experience	0.007 (0.99)	- 0.028 (0.93)
Workload (NASA TLX overall)	**0.856 (<0.01)**	- 0.009 (0.98)
Workload (NASA TLX MD)	0.533 (0.17)	- 0.184 (0.59)
Workload (NASA TLX TD)	**0.730 (0.04)**	- 0.198 (0.56)
Workload (NASA TLX effort)	**0.902 (<0.01)**	0.165 (0.63)
Workload (NASA TLX frustration)	0.011 (0.98)	0.287 (0.39)
Performance (GPS)	**0.758 (0.03)**	0.551 (0.08)
Performance (TAP)	- 0.688 (0.07)	- 0.396 (0.23)
ANTS global score		
Experience	0.057 (0.89)	0.146 (0.67)
Workload (NASA TLX overall)	**0.849 (<0.01)**	- 0.108 (0.75)
Workload (NASA TLX MD)	0.425 (0.29)	0.047 (0.89)
Workload (NASA TLX TD)	0.499 (0.21)	- 0.169 (0.62)
Workload (NASA TLX effort)	**0.927 (<0.01)**	- 0.080 (0.82)
Workload (NASA TLX frustration)	0.264 (0.53)	0.103 (0.76)
Performance (GPS)	0.504 (0.20)	**0.660 (0.03)**
Performance (TAP)	- 0.624 (0.10)	- 0.281 (0.40)

Table 9-8: Results of analysis of correlation (Pearson's correlation coefficient) between global and SA scores for ANTS with Nasa-TLX experience, performance and workload domains in whole teams

The positive correlation of ANTS (either the SA category or the global score) with workload is interesting, and may be because as workload rose, there were more obvious instances of the team members communicating their thoughts to each other.

It is possible that positive correlation of the ANTS global score with performance in scenario B is because other related constructs came to the fore such as task management: it was a busier scenario with more tasks to complete.

9.4 DISCUSSION

Other groups have reported correlations between SA and participant experience,[136,410] technical performance[136,409,410,438] and workload measures[137] in healthcare settings but none have studied real teams of AICU staff.

The previous chapter revealed the validity of one measure of SA (SAGAT) as a tool for discriminating between SA levels in easier and more complex scenarios for teams in AICU, evidence that was lacking for the observer based tool, ANTS. This final chapter considered the secondary outcome measures for the SASi study which focused primarily on assessing the relationship between measures of SA and experience, workload and performance. The discussion will now be divided into sections considering:

- Evidence of the reliability and validity of the measure of workload (NASA-TLX) in the SASi study
- Evidence of the validity (GPS and TAPs) and reliability (GPS) of the measures of performance in the SASi study
- The relationship between measures of SA and
    - Experience
    - Workload (NASA-TLX)
    - Performance (GPS and TAPs)

9.4.1 VALIDITY AND RELIABILITY OF NASA-TLX SYSTEM IN SASI

The NASA-TLX has been extensively validated in various work settings including healthcare and results for internal consistency have varied[439].

Evidence of discriminant validity for the NASA-TLX scores was shown both between and within scenarios.

9.4.1.1 ANALYSIS OF DIFFERENCES BETWEEN SCENARIOS

Significantly higher scores were found for workload (overall, in the temporal demand and frustration domains for doctors) in scenario B (the more complex scenario) than scenario A. The doctors in scenario B had, on average, 2.3 years more experience than those in scenario A (a factor that could not be controlled for in this study). Experience has been shown to correlate inversely with workload in doctors[451] but not nurses [452,453] i.e. the greater the experience the lower the perceived workload and this may, in part, explain why more obvious differences were not observed between scenario A and B for doctors (and, by extension, whole teams) in this study.

The story for the nurses was different in that there was no difference in NASA-TLX scores between the scenarios for either junior or senior nurses. This may be because the scenarios were not sufficiently separable in terms of workload for the nurses (although they were designed with the assistance of two senior AICU nurses). A study of 757 ICU nurses [452] found internal consistency (Cronbach's alpha) of the overall workload score (results for the sub-domains were not reported) to be 0.72 – similar to the result in the SASi study (0.69) before removal of the physical demand and performance domains. They also reported a similar lack of discrimination for experience level (which they defined by grouping the nurses into age groups).

Workload has been positively correlated with number and complexity of tasks undertaken[106,454] and for the nurses in both scenarios in the SASi study many of the tasks were similar (i.e. involved drawing up and administration of drugs and fluid, preparation of equipment for cardiac arrest etc). An analysis of the video data to count such tasks for the professional groups was not undertaken but may reveal additional insights into why workload differences were not observed for the nurses.

9.4.1.2 ANALYSIS OF DIFFERENCES WITHIN SCENARIOS

The goal directed task analyses for both scenarios (see Chapter 7) revealed that numbers of tasks and decisions were 60% greater in scenario B and increased as time went on. The analysis of workload at the three pause points in both scenarios revealed a significant increase from the beginning to the end for the whole team and professional groups.

Whilst the data distinguishing between the scenarios are more obvious for the doctors than the nurses the rate and extent of change within scenario B are similar for all professions providing evidence that the additional complexity was measurable with the NASA-TLX in all groups.

9.4.2 EVIDENCE FOR THE VALIDITY AND RELIABILITY OF PERFORMANCE MEASURES

9.4.2.1 GLOBAL PERFORMANCE SCORE (GPS)

The GPS was unable to discriminate between the team performances in Scenario A or B. None of the teams scored less than 50/100 and average scores were near 70/100 for both scenarios. Whilst this shows that scoring system we used was not hampered either by a floor or ceiling effect it highlights the fact that when measuring high performing teams, finding a measurement that allows discrimination between them is challenging. The lack of substantial difference

between teams may explain, in part, why other measures used in SASi could not discriminate clearly between teams or scenarios.

Reliability was good for the GPS when used by the two expert raters (HH and LV) in SASi and parallels results for other global scoring systems.[443] This is useful in the context of understanding the psychometric properties of the tool as it implies that it could be used in future SASi studies applied to theatre or ED settings.

9.4.2.2 TOTAL ACTION PERIODS (TAPS)

Scenario A took significantly longer to complete both in terms of length of the whole scenario and length of the standardised TAP than scenario B. We did not explicitly ask about perceived differences in the scenarios on the feedback form but during the debriefing sessions it was evident that participants felt, overall, there was less urgency in scenario A. The heightened sense of urgency in scenario B, particularly when the patient develops a tension pneumothorax and torsades de pointes, may have contributed to the decrease in TAPs evident in the recordings. The notion of using time taken to complete tasks as a measure of performance is not unusual in research in healthcare education. Two studies have compared SAGAT scores with performance times in healthcare. In the study by Morgan et al[150] a similar approach to the measurement of key time points was taken as in the SASi study and they found no significant difference in SAGAT scores between teams who completed tasks quickly and those who did not. The study by Zhang et al[148] compared SAGAT scores with time taken to recognise events on different types of display screen (traditional 2-D and graphical 3-D) in anaesthetists and bioengineering students. They found no difference between interfaces for anaesthetists but a significant reduction in detection times for the bioengineering students.

However, speed of completion of tasks may be more dependent on familiarity with the task itself or level of competency of the person undertaking it. All our teams were used to working together and none of the tasks was unfamiliar or uncommon in real ICU settings. It is likely that measuring time to completion was not a nuanced enough tool to measure performance in this context.

9.4.3 RELATIONSHIP OF MEASURES OF SA WITH LEVEL OF EXPERIENCE, WORKLOAD AND PERFORMANCE

9.4.3.1 EXPERIENCE

There was no correlation between experience and SA (either using SAGAT or ANTS), performance or workload in any of the professional groups or in teams. It is possible that higher levels of experience in doctors leading the teams in scenario B when SAGAT was used impacted SAGAT scores i.e. if less experienced doctors had been leading the teams in SAGAT scenario B we may have seen a greater difference in scores between the scenarios). A similar problem was observed in a study of workload in anaesthetists during real clinical activities Gaba et al[455] were unable to measure an impact of experience on workload as the more experienced anaesthetists were automatically allocated the more complicated cases.

A study of groups of medical students or doctors with different levels of experience (junior trainees, senior trainees and consultants) in simulated trauma scenarios revealed significantly higher SAGAT scores for more experienced groups[136] although this was most obvious at the extremes (i.e. between medical students and consultants). This study used artificially constructed, single discipline teams who would not normally work together in trauma settings

and it was unclear how many questions were asked at each stop and what proportion represented each level of SA.

Another study of trauma teams undertaking simulated trauma scenarios revealed significant increases in SAGAT scores between teams comprised of two medical and one nursing student and teams comprised of two senior trainees and a senior nurse or two consultants and a senior nurse.[410] However the teams were constructed specifically based on experience level, SAGAT questions were different depending on professional background, only 8 scenarios were studied (two for each of the four teams) and the differences were only obvious at extremes of experience as in the study by Hogan et al.[136] Conversely, and in line with the findings in the SASi study, McKenna et al[409] found no correlation between experience and SAGAT score in their study of 97 nursing students undertaking three simulation scenarios (mostly in teams of three) involving rapidly deteriorating patients. However, these teams were not "real" and only involved nurses. The SASi study involved teams of real AICU staff and was, therefore, more akin to a real clinical setting.

9.4.3.2 WORKLOAD: NASA-TLX

Results from previous studies correlating the NASA-TLX with SA have been positive, negative or equivocal.[439] The SASi study is the first in healthcare to analyse the association between SAGAT and workload using the NASA-TLX and no correlation was found. This may be due to the type of questions asked in SAGAT, to the lack of internal consistency found with the NASA-TLX scale in the SASI study (until two of the domains were removed) or to the inconsistencies in experience within the teams.

In a study undertaken in a real emergency department[137] workload was measured in doctors using number of "work events" (e.g. arranging to admit a patient to hospital, discharging a patient, ordering a test or a medication), number of patients managed and calculation of a weighted workload score. Lower SA scores were found in doctors with more patients in their care. In the SASi scenarios there was only one patient but in real AICU settings the doctors would be responsible for many more patients at a time. Future research could include standardised scenarios where more than one patient was involved to mimic real life more accurately.

9.4.3.3 MEASURES OF PERFORMANCE (GPS AND TAPS)

The primary outcome measure for SASi was to consider the validity and reliability of measures of SA in simulated AICU scenarios. This included an assessment of correlations with other related constructs and other studies have analysed the relationship of SAGAT with experience and performance with differing results.

Two studies in teams undertaking trauma scenarios[136,419] found evidence of a positive correlation for SAGAT with a trauma checklist. Both studies also found a positive correlation with experience and SAGAT score but results were only significant for differences between students or junior residents and consultants and no account was given of how experience was quantified.

In a study of 43 third year medical students working in ten teams of four or five, in two trauma scenarios, evidence of validity was found through a significant positive correlation of SAGAT score with the Mayo High Performance Teamwork Scale. However, there were only nine questions per scenario and half points were allocated for "questions deserving partial credit"

which included correctly identifying a tension pneumothorax but on the wrong side – an error which could cause significant harm to the patient if not corrected.

SAGAT has also been studied in nursing students. Cooper et al[438] studied 51 final year nursing students using two short scenarios (one on septic shock and one on hypovolaemia) and there was no correlation between SAGAT score, multiple choice questionnaires testing knowledge or technical performance. McKenna et al[409] found that SAGAT scores in 97 final year nursing students were low (mean score 41% with the lowest scores attributed to perception at 26%) and were only significantly positively associated with technical performance in one of the three emergency scenarios. However scenarios were very short (approximately seven minutes) and the 12 item SAGAT questionnaire was used at the end of the scenario rather than using pauses.

9.5 STUDY LIMITATIONS

Whilst this is the largest study of its kind (using real teams of healthcare professionals) resource limitations meant that SAGAT questionnaires could only be applied to one scenario. Data for our analysis of SAGAT in full teams, therefore, were limited to 8 for scenario A and 11 for scenario B.

The NASA-TLX was answered at the same time as the SAGAT questions and we did not make adequate adjustments for thinking time for the candidates which led to the pause times being longer than recommended. Results from the questionnaire, however, revealed that the majority of participants did not feel that the pauses disrupted their learning.

It was not possible to control for experience levels in the two scenarios as staff attending training were chosen by the education team on the day and were allocated according to acuity

of case load on the AICU. The doctors in scenario B were significantly more experienced than those in scenario A (the easier scenario).

As described in the results, a significant proportion of staff attending the training had English as a second language and the nuances of translating questions on SA or workload may have led to inadvertent misunderstandings – these were mitigated by giving careful instruction on the use of each questionnaire and by having a member of faculty available for each participant.

9.6 CONCLUSION AND FUTURE RESEARCH

The SASi study has found evidence of discriminant validity for the NASA-TLX score in two simulated scenarios of AICU emergencies. Correlations were found between the observer-based measure of SA (ANTS) and GPS and workload but these did not exist for SAGAT. Review of the literature has found conflicting results. The lack of standardisation of methodologies for the design and delivery of SAGAT, variability in participants studied, measures of performance and definitions of validity as well as techniques used for data analysis hamper any meaningful interpretation of the available evidence for the use of SAGAT in healthcare. This study has highlighted a number of important unanswered questions about the measurement of SA in simulated settings and these will be considered in the final chapter.

Future research using SAGAT should seek to provide a standard methodology for SAGAT development (including banks of scenarios for use across research communities) and a consensus on measures of workload and performance to use in furthering our understanding of SA in healthcare.

CHAPTER 10 DISCUSSION AND CONCLUSIONS

"I see no more than you, but I have trained myself to notice what I see."

Sherlock Holmes in: The Adventure of the Blanched Soldier[456]

A recent review of SA has highlighted that it is one of the most widely studied and hotly debated topics in ergonomics[94] but SA is still a novel concept in medicine. The NHS is 70 years old and it is only in the past two decades that NTS (SA included) have emerged as important aspects of high quality clinical practice and only in the last ten years that they've been embedded in the curriculum of my own profession of anaesthesia. This thesis has come about because I wanted to draw attention to the importance of SA (and other NTS) in clinical and educational contexts in healthcare. Whilst this work has provided greater clarity on some aspects of SA there remain many unanswered questions which I will highlight in this final chapter as they provide opportunities for exciting research in the future.

This chapter will be divided into the following sections:

- A summary of the findings of the thesis
- A discussion of the strengths and weaknesses of the methods
- The implications for training in healthcare
- Opportunities for future research and finally
- Wider policy implications

10.1 SUMMARY OF FINDINGS

The studies in this thesis have explored SA error in serious incidents in the OUHT; analysed the tools available to measure NTS (including SA); compared tools for the assessment of NTS in

simulated cardiac arrest situations and investigated techniques for the measurement of SA in simulation training for AICU staff. I will now summarise the findings.

The design and implementation of an holistic method of review of SIRIs in the OUHT in Chapters 3 and 4 found evidence of SA errors in 96% of the most serious incidents in the organisation. This finding resonated with other studies in healthcare[78,79,99] but when considered in the context of the incident attributes and error types a deeper understanding of the role of SA was evident:

- The majority of SA errors happened at level 1 (perception)
- Incidents that happened over a short time frame (acute SIRIs) were found to be more likely to involve SA errors at the level of comprehension (i.e. sense-making) and projection (i.e. prediction of developments in the near future)
- Knowledge based errors were more likely in acute SIRIs
- There were problems with the use of standard operating procedures (SOPs) in 89% of all SIRIs (86% of acute SIRIs) which impacted SA
- Distraction was a contributory factor in 41% of acute SIRIs
- Issues with equipment or technology were a factor in 44% of all SIRIs and 59% of acute SIRIs

This list of key findings in the thematic analysis highlights the value of an holistic approach to incident analysis. Too often in healthcare the focus of the investigation is narrow. The value of using a "human factors lens" through which to view critical incidents is in the strength of the recommendations that result. For example these data provide insights into the use of SOPs (or lack of) in the OUHT and, specifically, the clinical areas where the acute SIRIs are happening. This has implications for governance teams in the Trust (e.g. for the design and implementation

of SOPs through engagement with staff groups that should be using them) and for intelligent targeting of team training programmes for frontline staff to help them cope more effectively with acutely deteriorating patients.

The analysis of the SIRIs in Chapter 4 highlighted the need to understand how SA is measured in healthcare and the key findings from the systematic review in Chapter 5 were:

- In most cases SA is measured as one of a group of NTS, using an observer-based tool
- There is a bewildering array of tools for the assessment of NTS in healthcare
- The evidence of validity and reliability of the tools is very variable

The ideal tool for NTS assessment in healthcare does not yet exist. Further research is required to determine if a more generic tool for use in any healthcare context is feasible.

The study of expert raters using four different tools for NTS assessment in Chapter 6 revealed some interesting and unexpected findings:

- Despite lengthy experience in the formative assessment of NTS in simulation three expert raters found the use of unfamiliar but well validated tools for NTS assessment challenging
- Internal consistency was poorest for the tool we were most familiar with (ANTS)
- Interrater reliability was surprisingly poor

These data reinforced the results of the systematic review in Chapter 5 regarding variability in method of design and usability but clearly highlighted the limited use of appropriate statistical analyses to measure reliability of the tools. A key question to consider when using NTS assessment is “how much reliability is enough?” and the answer will be different dependent on the context of use. The study highlighted the challenges faced in providing consistency in NTS

assessments and the clear need for the presence of more than one rater in high stakes settings where excellent reliability is essential.

The final three chapters (7, 8 and 9) were studies of simulation training for AICU teams (the SASi study) which built on the themes emerging from the earlier work. I was interested in understanding if there was a better way to measure the invisible construct of SA with a view to developing targeted training interventions similar to those used in military settings. Three measures of SA were examined and key findings were as follows:

- SAGAT was used for the first time in AICU teams and was able to discriminate more effectively between scenarios and staff but required considerable effort to construct
- Evidence of validity for ANTS was limited – the scores did not discriminate between scenarios
- Evidence of validity for the subjective measure of SA (SART) was poor and in the context of the SASi study it is unlikely it was measuring SA

The SASi study also analysed the relationships between measures of SA and experience, workload and performance. No correlations were found for SAGAT but for the ANTS SA category there were positive correlations with workload and performance in scenario A but only for the ANTS global score and performance in scenario B. These results raise some obvious questions about the measurement of SA and the SASi study has laid the groundwork for using similar, standardised scenarios in studies of SA in other acute care settings such as ED and operating theatres.

10.2 STRENGTHS AND LIMITATIONS OF METHODS

All the methods used in this thesis have been designed with standardisation of techniques, harmonisation of data analysis and reduction of bias in mind, the strengths and limitations will be explored in this next section by relevant chapters.

Chapters 3 and 4 (SIRI analysis):

A novel, holistic method of incident analysis was devised and used for the study of 167 serious incidents in the OUHT. This method combined a human factors approach with definitions of error type and a description of the role of SA in each incident and provided valuable insights into a cohort of cases which developed over a shorter time frame. The analysis of serious incidents in the OUHT was a sample of convenience i.e. all 167 incident reports from 2015-16 were included. Some of the studies referenced in this thesis have analysed thousands of reports but this was the first year in which a more rigorous screening of incidents had been undertaken in the OUHT alongside the initiation of the SIRI forum. The method developed for the holistic analysis of critical incidents will require further validation on future cohorts of incidents. The analysis also highlighted the variability in quality of the reports which is a problem encountered in other reviews of incidents using post hoc analysis, however, this variability was mitigated by use of the same report framework for every incident.

Chapter 5 (systematic review)

A deliberately pragmatic approach was adopted in the systematic review of tools for NTS assessment in healthcare. I decided that a producing a list of the attributes of the 76 tools would provide a comprehensive picture of those available but would not be any assistance in choosing the best tool for a given setting. A novel scoring system was, therefore, devised to

rank the tools according to method of development and rigour of psychometric testing. The scoring system used to rank the tools, however, had some limitations:

- It favoured tools that were for use in a wide variety of healthcare settings, not just one
- It favoured tools for more than one discipline
- Tools which were only recently developed lost marks if they had not yet provided evidence of validity
- Older tools lost marks because they had nothing to compare with and their methods of development were not as robust as those which benefitted from the more recent work on psychometric testing and the importance of a scientific approach to the identification of relevant NTS

This system was devised using the best available evidence for the design of such tools and evidence of its own reliability was good. Furthermore, there are no reviews of assessment tools for either technical or NTS in which a decision tree has been produced to provide a useful reference for educators in healthcare.

Chapter 6 (NTS assessment of standardised, simulated cardiac arrest videos)

This study used four tools for the assessment of NTS in simulated cardiac arrests and revealed that expert raters may not provide consistent scores. Thus, a potential risk was exposed if isolated raters use such tools in high stakes settings. Although it may have benefitted from higher numbers of videos scored other studies have used similar numbers to provide evidence of reliability. The choice of the tools was guided by the use of the decision tree developed from the systematic review and yet OTAS, despite its very good score, was quickly rejected as a viable tool from the standpoint of usability. Similarly, OSCAR, whilst it scored highly for

reliability, scored poorly on the assessment of usability. This has revealed that evidence of validity and reliability do not tell the whole story for these methods of assessment and that usability is an important factor in deciding on the most appropriate tool in a given setting.

Chapters 7,8 and 9 (SASi study)

One of the themes throughout this thesis has been the lack of standardisation of methods found in other studies (i.e. for incident analysis and design and testing of tools for the assessment of NTS in healthcare) and this area of educational research was no exception. The methods in the SASi study, therefore, involved the standardisation of programming of scenarios such that the same physiological response would be experienced by the teams in the scenarios for every training session e.g. blood pressure would change at the same rate and respond to treatment in the same way and treatment of the pneumothorax would result in the same improvement of oxygen saturation. Furthermore, every session was run by the same team of faculty and technicians to reinforce the consistency of delivery of the scenarios. The problems faced in delivering the training, however, were exactly those we face day-to-day in the NHS, namely the difficulty of releasing staff in the context of a busy clinical unit. This was partially overcome by the excellent relationship with the education and management team in AICU who worked very hard to ensure teams were released. The result was that timing constraints only allowed one SAGAT scenario per session and the random allocation meant experience levels could not be controlled for in the SAGAT scenarios. However, this was a pragmatic study of real teams and the first and largest of its kind in AICU and it provided evidence of the significant challenge of designing SAGAT scenarios and, more generally, the difficulties of measuring SA.

The three tools used to measure SA only considered SA at individual level and not at team or system levels. Methods have been described to calculate team SA based on individual SAGAT

scores but this did not provide evidence of validity for SAGAT in the SASi study. Furthermore, these statistical techniques do not provide the qualitative evidence of how shared SA is attained e.g. evidence of the quality or content of verbal or non-verbal communication between team members.

10.3 IMPLICATIONS FOR TRAINING

No healthcare professional works in isolation but teams are formed on a much more ad hoc basis in clinical settings than was the case before the advent of the European Working Time Directive. These changes have brought home the fact that good teamwork cannot be seen as simply a natural product of working together with familiar colleagues, but must be embedded in the wider healthcare system. Multidisciplinary healthcare professionals are expected, in both elective and (more commonly) emergency settings, to perform with seamless efficiency in the context of having little knowledge of the skills or competencies of the colleagues they will be working with. If we are to continue to accept this situation as the norm and rely on serendipity rather than resilience to ensure high standards of care then avoidable harm rates will surely remain as they are.

Part of the reason for the lack of progress in safety in healthcare lies in the way we train our healthcare professionals, specifically, we do not offer them regular opportunities to practise crisis situations in simulated settings where teams might make mistakes without the risk of harm to the patient. The immersive simulation training which is currently offered to support the development of these team competencies is provided on an ad hoc basis and is neither standardised nor quality assured.

In Chapter 2, I described the training interventions which have been used to improve SA in work settings outside healthcare and the principles would, ostensibly, be transferrable to clinical areas. Before designing and implementing SA targeted training, however, there are several unanswered questions revealed by the studies in this thesis regarding the measurement of SA:

- How do we rationalise the measurement of NTS (including SA) in healthcare and can educators be better supported in using tools for the assessment of NTS?
- What does the use of a direct, objective measure of SA (such as SAGAT) add to our understanding of SA in simulated clinical environments?
- Can SAGAT be adapted to measure team SA in simulated settings more effectively or, do we need a different tool?
- What is the correlation between SA experience, workload and performance and how do we measure it?

There is evidence to support the use of team training incorporating simulation to improve performance and outcome in healthcare but training interventions targeting SA have yet to be designed and tested for clinical settings. There should be a focus on developing scenarios which allow training in, for example, improved visual search strategies to highlight the importance of good team SA. Results from this thesis would suggest that ensuring we have a valid and reliable tool for the measurement of SA remains a priority alongside the development of these novel training interventions.

10.4 FUTURE RESEARCH

SA error has been shown to be a significant factor in the evolution of adverse events in the OUHT. Future research should include extended use of the holistic method of incident analysis

described in Chapter 3 for thematic analysis of incidents occurring in the OUHT and the wider NHS (including primary care settings). Additional human factors training for incident analysis in the OUHT has already begun but automation of data input and intelligent targeting of investigations for "near miss" events in the organisation would rationalise the workload and improve the quality of reports overall. More importantly understanding the cause of errors occurring in the Trust would direct recommendations for prevention of further incidents that are more likely to be successful and sustainable. The London Protocol has been underutilised in the NHS, yet it provided the framework for a much richer understanding of the context of SA error, particularly in the acute SIRIs in this thesis. Future research should incorporate analysis of the structured interview process in the protocol. This was designed to support investigators in capturing important aspects of behaviour and systems problems which play a role in incidents and could improve the objectivity, consistency and thoroughness of the investigation.

Measurement of SA has repeatedly been shown to be challenging (see Chapters 6-9) in this thesis. There are too many tools available for the measurement of NTS in healthcare and evidence from the video review chapter highlighted that evidence of validity and reliability do not tell the whole story when one is choosing a tool to use for the assessment of NTS. This is an ideal moment to bring together a working group of subject matter experts to consider how best to rationalise the situation and provide clear guidance for educators in healthcare.

The self-assessment tool for the measurement of SA (SART) was not found to be reliable in the context of the SASi study and reinforced the problems faced in using self-assessment tools as useful measures of SA or other NTS. There was evidence, however, that SAGAT was able to discriminate between an easy and a more difficult scenario and between professional groups. The problem lies in the time taken to design a questionnaire for SAGAT (it took 10 subject

matter experts two months for two scenarios in the SASi study) and that it does not measure team SA. In fact none of the available measures was wholly suitable for the measurement of team SA in the SASi study. More empirical evidence is required to define the ideal measure, or combination of measures, of SA in individuals, teams and wider systems and to provide greater clarity on its relationship with team performance in healthcare.[94]

It seems unlikely that SA is unrelated to workload or performance of teams in healthcare. The lack of clear evidence from the SASi study is more likely due to the measures used than to an absence of correlation. Future research should examine more closely the relationship between teams of, for example, highly experienced staff and novices to tease apart which factors lead to good SA and which factors are barriers. Whilst proof of concept and standardisation of "stressors" is more straightforward in a simulation room, studies will be necessary in the real clinical workplace to understand how we best measure and train for good SA in healthcare teams.

10.5 WIDER POLICY IMPLICATIONS

Despite the extensive literature on error in healthcare, strategies to implement sustainable change[457] and develop safety training in the NHS, [458] [459] the pace of change is slow.[460] Several important reports have recently been published and highlight the considerable challenges faced by healthcare professionals and NHS managers in delivering safe, high quality care.

The most recent GMC national training survey is a sobering read and key conclusions from the Training Environments 2018[461] report are as follows:

- Heavy, intense workloads can disrupt training, and sometimes lead to doctors in training working beyond their competence or experience

- Poor handovers and inductions can have a negative impact on trainees' education and development, and can lead to issues with continuity of care for patients
- The lack of time to train remains a significant issue for many trainers
- Intense workloads and rota gaps often disrupt training
- A quarter of doctors in training and a fifth of trainers reported feeling burnt out to a high or very high degree

The authors of the report conclude that:

"Looking to the future, we believe there needs to be a greater focus on the essential human factors that underpin professional behaviour, promote safe and effective practice, wellbeing and foster a positive organisational culture…..That's why we have prioritised human factors in the Generic Capabilities Framework. All colleges and faculty curricula must now show how human factors have been integrated into specialty training."

These aspirations are laudable but a recent review of the healthcare workforce by the Health Foundation, the King's Fund and the Nuffield Trust[462] revealed that "the workforce challenges in the NHS in England now present a greater threat to health services than the funding challenges" and this view was reinforced in Lord Darzi's report, "Better Health and Care for All" in which he states, " a properly funded NHS is the foundation on which a fair, cohesive and inclusive society Is built" but highlights that 11% of nursing posts, 12% of GP posts and 5% of medical consultant posts are unfilled.[463] Against this backdrop of inadequate staffing and the recognition that the NHS has been subject to cycles of feast and famine in funding, the report recommends that funding to the NHS should be increased but that it should be provided more efficiently such that long term investments are possible rather than short-term fixes. The past two decades of increasingly obvious research and public scrutiny into patient safety provide the

impetus, despite funding constraints, for the development of sustainable team training as an essential requirement of a safe system.

10.6 FINAL REFLECTIONS

It remains to be seen whether recommendations from these reports impact government spending on the NHS and in the meantime the pressing issue of improving patient safety remains. Some of the solutions already exist in the OUHT and have been highlighted in this thesis. Using thematic analysis to direct incident investigation, provide robust recommendations for change and intelligently target training interventions could reduce inefficiency in the governance teams and use the limited simulation training resource more effectively. These interventions take time to deliver results, however, and require research that sits alongside them to monitor outcomes and inform next steps. My personal goal is to ensure that human factors training and research is embedded in the OUHT over my career and beyond.

REFERENCES:

1. Berwick D. *A Promise to Learn – a Commitment to Act Improving the Safety of Patients in England*.; 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/226703/Berwick_Report.pdf. Accessed March 9, 2019

2. Beecher H, Todd D. A Study of the Deaths Associated with Anaesthesia and Surgery. *Ann Surg*. 1954;140:2-35.

3. Dripps RD, Lamont A, Eckenhoff JE. The Role of Anesthesia in Surgical Mortality. *JAMA*. 1961;178(3):261-266.

4. Clifton BS, Hotten WIT. Deaths Associted With Anaesthesia. *Br J Anaesth*. 1963;35(4):250-259.

5. Cooper JB, Newbower RS, Long CD, McPeek B. Preventable Anesthesia Mishaps: A Study of Human Factors. *Anesthesiology*. 1978;49(6):399-406.

6. Williamson JA, Webb RK, Sellen A, Runciman WB, Van der Walt JH. The Australian Incident Monitoring Study. Human failure: an analysis of 2000 incident reports. *Anaesth Intensive Care*. 1993;21(5):678-683.

7. American Society of Anesthesiologists Closed Claim Project. http://depts.washington.edu/asaccp/about-us/anesthesia-quality-institute. Accessed March 9, 2019

8. Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Eng J Med*. 1991;324(6):370-376.

9. Bhasale AL, Miller GC, Reid SE, Britt HC. Analysing potential harm in Australian general practice: an incident-monitoring study. *Med J Aust.* 1998;169(2):73-76.

10. Beckmann U, Baldwin I, Hart GK, Runciman WB. The Australian Incident Monitoring Study in Intensive Care: AIMS-ICU. An analysis of the first year of reporting. *Anaesth Intensive Care*. 1996;24(3):320-329.

11. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ*. 2001;322(7285):517-519.

12. Makary M, Daniel M. Medical error—the third leading cause of death in the US. *BMJ*.

2016;353(i2139).

13. Schiøler T, Lipczak H, Pedersen BL, et al. [Incidence of adverse events in hospitals. A retrospective study of medical records]. *Ugeskr Laeger*. 2001;163(39):5370-5378. http://www.ncbi.nlm.nih.gov/pubmed/11590953.

14. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals I: occurrence and impact. *N Z Med J*. 2002;115(1167):U271.

15. Baker GR, Norton PG, Flintoft V, et al. The Canadian Adverse Events Study: The incidence of adverse events among hospital patients in Canada. *CMAJ*. 2004;170(11):1678-1686.

16. Kohn LT, Corrigan JM, Molla S. *To Err Is Human. Building a Safer Healthcare System*. Washington DC: National Academy Press; 1999.

17. Officer CM. *An Organisation with a Memory*.; 2001. https://www.aagbi.org/sites/default/files/An organisation with a memory.pdf. Accessed March 9, 2019.

18. Elaine Bromiley's Story. http://s753619566.websitehome.co.uk/wp-content/uploads/2018/06/ElaineBromileyAnonymousReport.pdf. Accessed March 2019

19. National Quality Board. Human Factors in Healthcare - A Concordat from the National Quality Board. 2013. https://www.england.nhs.uk/wp-content/uploads/2013/11/nqb-hum-fact-concord.pdf. Accessed March 9, 2019

20. Roberts KH. Managing High Reliability Organizations. *Calif Manage Rev*. 1990;32(4):101-114.

21. Reason JT. *Human Error*. Cambridge University Press; 1990.

22. Reason J. Understanding adverse events: human factors. *Qual Health Care*. 1995;4(2):80-89.

23. Perrow C. *Normal Accidents : Living with High-Risk Technologies*. Princeton University Press; 1999.

24. Amalberti R. The paradoxes of almost totally safe transportation systems. *Saf Sci*. 2001;37(2-3):109-126.

25. Ferner RE. An agenda for UK clinical pharmacology medication errors. *Br J Clin Pharmacol*. 2012;73(6):912-916.

26. Shore DA. *The Trust Crisis in Healthcare : Causes, Consequences, and Cures*. Oxford University Press; 2007.

27. Gibbs N, Borton C. *Safety of Anaesthesia in Australia A Review of Anaesthesia Related Mortality 2000 - 2002*.; 2006. http://www.anzca.edu.au/documents/safety-of-anaesthesia-in-australia-2000-2002.pdf. Accessed March 9 2019.

28. Morey JC, Simon R, Jay GD, et al. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. *Health Serv Res*. 2002;37(6):1553-1581.

29. RW F, JM W, Torrie J, et al. The effect of a simulation-based training intervention on the performance of established critical care unit teams. *Crit Care Med*. 2011;39(12):2605-2611.

30. Schouten LMT, Hulscher MEJL, van Everdingen JJE, Huijsman R, Grol RPTM. Evidence for the impact of quality improvement collaboratives: systematic review. *BMJ*. 2008;336(7659):1491-1494.

31. Weaver SJ, Dy SM, Rosen MA. Team-training in healthcare: a narrative synthesis of the literature. *BMJ Qual Saf*. 2014;23(5):359-372.

32. Health and Safety Executive. Introduction to Human Factors. http://www.hse.gov.uk/humanfactors/introduction.htm. Accessed March 9, 2019

33. Reason JT. *The Human Contribution : Unsafe Acts, Accidents and Heroic Recoveries*. Ashgate; 2008.

34. Reason J. Human error: models and management. *BMJ*. 2000;320(7237):768-770.

35. Singh H, Thomas E, Petersen L, Studdert D. Medical Errors Involving Trainees. *Arch Intern Med*. 2007;167(19):2030-2036.

36. Lavie N. The role of perceptual load in visual awareness. *Brain Res*. 2006;1080(1):91-100.

37. Lavie N, Lin Z, Zokaei N, Thoma V. The role of perceptual load in object recognition. *J Exp Psychol Hum Percept Perform*. 2009;35(5):1346-1358.

38. Kahneman D. *Attention and Effort*. First. New Jersey: Prentice-Hall; 1973.

39. Kahneman D. *Thinking, Fast and Slow*. Allen Lane; 2011.

40. Pew RW. The State of Situation Awareness Measurement: Heading Toward the next Century. In: Endsley MR, Garland DJ, eds. *Situation Awareness Analysis and Measurement.* Mahwah, New Jersey: Taylor & Francis; 2000:33-47.

41. Graham ER, Burke DM. Aging increases inattentional blindness to the gorilla in our midst. *Psychol Aging*. 2011;26(1):162-166.

42. Becker MW, Leinenger M. Attentional Selection Is Biased Toward Mood-Congruent Stimuli. 2011.

43. Chabris CF, Simons DJ. *The Invisible Gorilla : And Other Ways Our Intuitions Deceive Us*. Crown; 2010.

44. Cheyne JA, Carriere JSA, Smilek D. Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Conscious Cogn*. 2006;15(3):578-592.

45. Cowan N. What are the differences between long-term, short-term, and working memory? *Prog Brain Res*. 2008;169:323-338.

46. Flin R, O'Connor P, Crichton M. Safety at the sharp end: A guide to non-technical skills; Chapter 2. In: *Safety at The Sharp End: A Guide to Non-Technical Skills*. ; 2008:17-40.

47. Miller GA. The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychol Rev*. 1955;101(2):343-352.

48. Atkinson RC, Shiffrin RM. Human Memory: A Proposed System and its Control Processes. In: Spence K, Spence J, eds. *The Psychology of Learning and Motivation: Advances in Research and Theory*. New York: Academic Press; 1968:89-195.

49. Baddeley A. Working memory: looking back and looking forward. *Nat Rev Neurosci*. 2003;4(10):829-839.

50. Shah P, Miyake A. Models of working memory - an introduction. In: Shah P, Miyake A, eds. *Models of Working Memory - Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press; 1999:1-27.

51. Gruszka A, Nęcka E. Limitations of working memory capacity: The cognitive and social consequences. *Eur Manag J*. 2017;35(6):776-784.

52. Fabiani M. It was the best of times, it was the worst of times: A psychophysiologist's view of cognitive aging. *Psychophysiology*. 2012;49(3):283-304.

53. Schoofs D, Preuss D, Wolf OT. Psychosocial stress induces working memory impairments in an n-back paradigm. *Psychoneuroendocrinology*. 2008;33(5):643-653.

54. Baddeley A. Working memory. *Science*. 1992;255(5044):556-559.

55. Bartlett FC, Burt C. Remembering: A Study in Experimental and Social Psychology. *Br J Educ Psychol*. 1933;3(2):187-192.

56. Zsambok CE, Klein G. *Naturalistic Decision Making.* Taylor and Francis; 2014.

57. Johnson-Laird. Mental models in cognitive science. *Cogn Sci*. 1980;4:71-115.

58. Johnson-Laird P. The History of Mental Models. In: Manktelow K, Chung M, eds. *Psychology of Reasoning: Theoretical and Historical Perspectives*. Psychology Press; 2004:179-212.

59. Reason J. Studies of Human Error. In: *Human Error*. Cambridge University Press; 1990:19-52.

60. Craik KJW. *The Nature of Explanation*. Cambridge University Press; 1943.

61. Brewer W. Schemas versus mental models in human memory. In: Morris P, ed. *Modelling Cognition*. Oxford, England: John Wiley & Sons; 1987:187-197.

62. Rouse W. Models of Human Problem Solving: Detection, Diagnosis, and Compensation for System Failures. In: *IFAC Conference on Analysis, Design and Evaluation of Man-Machine Systems*. Baden Baden; 1982:167-184.

63. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science (80-)*. 1974;185:1124-1131.

64. Rasmussen J, Jensen A. Mental Procedures in Real-Life Tasks: A Case Study of Electronic Trouble Shooting. *Ergonomics*. 1974;17(3):293-307.

65. Rasmussen J. Human errors. A taxonomy for describing human malfunction in industrial installations. *J Occup Accid*. 1982;4(2-4):311-333.

66. Nestel D, Walker K, Simon R, Aggarwal R, Andreatta P. Nontechnical Skills: an inaccurate and unhelpful descriptor? *Simul Healthc J Soc Simul Healthc*. 2011;6(1):2-3.

67. Gaba DM. Training and Nontechnical Skills: The Politics of Terminology. *Simul Healthc J Soc Simul Healthc*. 2011;6(1):8-10.

68. Flin R., O'Connor P, Crichton M. Safety at the Sharp End: A Guide to Non-technical Skills; Chapter 1. In: *Safety at the Sharp End*. ; 2008:1-16.

69. Klampfer B, Flin R, Helmreich RL, et al. *Enhancing Performance in High Risk Environments: Recommendations for the Use of Behavioural Markers*. Zurich; 2001. http://www.raes-hfg.com/reports/notechs-swiss.pdf. Accessed March 9, 2019.

70. Hull L, Arora S, Symons NRA, et al. Training Faculty in Nontechnical Skill Assessment National Guidelines on Program Requirements. *Ann Surg*. 2013;258:370-375.

71. Gordon M, Baker P, Catchpole K, Darbyshire D, Schocken D. Devising a consensus definition and framework for non-technical skills in healthcare to support educational design: A modified Delphi study. *Med Teach*. 2015;37:572-577.

72. National Transportation Safety Board. *Aviation Accident Report AAR-95-03*.; 1995. https://www.ntsb.gov/investigations/AccidentReports/Pages/AAR9503.aspx. Accessed March 9, 2019.

73. National Transportation Safety Board. *Aviation Accident Report AAR-88-05*.; 1988. https://www.ntsb.gov/investigations/AccidentReports/Pages/AAR8805.aspx. Accessed March 9, 2019.

74. International Nuclear Safety Advisory. *The Chernobyl Accident: Updating of INSAG-1*. Vienna; 1992. https://www-pub.iaea.org/MTCD/publications/PDF/Pub913e_web.pdf. Accessed March 9, 2019.

75. Singh H, Hirani K, Kadiyala H, et al. Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record-based study. *J Clin Oncol*. 2010;28(20):3307-3315.

76. Singh H, Petersen LA, Thomas EJ. Understanding diagnostic errors in medicine: a lesson from aviation. *Qual Saf Health Care*. 2006;15(3):159-164.

77. Singh H, Daci K, Petersen LA, et al. Missed opportunities to initiate endoscopic evaluation for colorectal cancer diagnosis. *Am J Gastroenterol*. 2009;104(10):2543-2554.

78. Schulz CM, Krautheim V, Hackemann A, Kreuzer M, Kochs EF, Wagner KJ. Situation awareness errors in anesthesia and critical care in 200 cases of a critical incident reporting system. *BMC Anesthesiol*. 2016;16(1):4.

79. Schulz CM, Burden A, Posner KL, et al. Frequency and Type of Situational Awareness Errors Contributing to Death and Brain Damage. *Anesthesiology*. 2017;127:326-337.

80. Endsley MR. Design and Evaluation for Situation Awareness Enhancement. In: *Proceeding of the Human Factors Society 32nd Annual Meeting*. ; 1988:97-101.

81. Spick M. *The Ace Factor : Air Combat and the Role of Situational Awareness*. Naval Institute Press; 1988.

82. Wiener EL, Curry RE. Flight-deck automation: promises and problems. *Ergonomics*. 1980;23(10):995-1011.

83. Fracker ML. A Theory of Situation Assessment: Implications for Measuring Situation Awareness. *Proc Hum Factors Soc Annu Meet*. 1988;32(2):102-106.

84. Sarter NB, Woods DD. Situation Awareness: A Critical But Ill-Defined Phenomenon. *Int J Aviat Psychol*. 1991;1(1):45-57.

85. Banbury S, Tremblay S. *A Cognitive Approach to Situation Awareness : Theory and Application*. Ashgate Pub; 2004.

86. Dekker S, Hollnagel E. Human factors and folk models. *Cogn Technol Work*. 2004;6(2):79-86.

87. Flach JM. Situation Awareness: Proceed with Caution. *Hum Factors*. 1995;37(1):149-157.

88. Uhlarik J, Comerford DA. A Review of Situation Awareness Literature Relevant to Pilot Surveillance Functions. 2002. https://www.skybrary.aero/bookshelf/books/239.pdf. Accessed March 9, 2019.

89. Salmon PM, Stanton NA, Walker GH, Jenkins D, Baber C, McMaster R. Representing situation awareness in collaborative systems: A case study in the energy distribution domain. *Ergonomics*. 2008;51(3):367-384.

90. Fioratou E, Flin R, Glavin R, et al. Beyond monitoring: Distributed situation awareness in anaesthesia. *Br J Anaesth*. 2010;105(1):83-90.

91. Klein G, Moon B, Hoffman RR. Making Sense of Sensemaking 1: Alternative Perspectives. 2006. http://perigeantechnologies.com/publications/MakingSenseofSensemaking1-AlternativePerspectives.pdf. Accessed March 9, 2019.

92. Chiappe DL, Strybel TZ, Vu K-PL. Mechanisms for the acquisition of situation awareness in situated agents Mechanisms for the acquisition of situation awareness in situated agents. *Theor Issues Ergon Sci*. 2012;13(6):625-647.

93. Liu S, Wanyan X, Zhuang D. Modeling the situation awareness by the analysis of cognitive process. *Biomed Mater Eng*. 2014;24(6):2311-2318.

94. Stanton NA, Salmon PM, Walker GH, Salas E, Hancock PA. State-of-science: situation awareness in individuals, teams and systems. *Ergonomics*. 2017;60(4):449-466.

95. Wickens CD. Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement. *Hum Factors*. 2008;50(3):397-403.

96. Gaba DM, Howard SK, Small SD. Situation awareness in anesthesiology. *Hum Factors*. 1995;37(1):20-31.

97. Nemeth CP. *Improving Healthcare Team Communication : Building on Lessons from Aviation and Aerospace*. Ashgate; 2008.

98. Pronovost P, Berenholtz S, Dorman T, Lipsett PA, Simmonds T, Haraden C. Improving communication in the ICU using daily goals. *J Crit Care*. 2003;18(2):71-75.

99. Reader TW, Flin R, Mearns K, Cuthbertson BH. Team situation awareness and the anticipation of patient progress during ICU rounds. *BMJ Qual Saf*. 2011;20(June):1-8.

100. Reader T, Flin R, Lauche K, Cuthbertson BH. Non-technical skills in the intensive care unit. *Br J Anaesth*. 2006;96(5):551-559.

101. Wellens A. Group situation awareness and distributed decision making: from military to civilian applications. In: Castellan NJ, Association. AP, eds. *Individual and Group Decision Making : Current Issues*. L. Erlbaum Associates; 1993:315.

102. Adams MJ, Tenney YJ, Pew RW. Situation Awareness and the Cognitive Management of Complex Systems. *Hum Factors*. 1995;37(1):85-104.

103. Klein G. Sources of Error in Naturalistic Decision Making Tasks. *Proc Hum Factors Ergon Soc Annu Meet* . 1993;37(4):368-371.

104. Endsley MR. Towards a Theory of Situation Awareness in Dynamic Systems. *Hum Factors*. 1995;37(1):32-64.

105. Endsley MR. Theoretical Underpinnings of Situation Awareness: a Critical Review. In: Endsley MR, Garland DJ, eds. *Situation Awareness Analysis and Measurement.* Mawah, New Jersey: CRC Press; 2000:3-33.

106. Patten CJ., Kircher A, Östlund J, Nilsson L. Using mobile telephones: cognitive workload and attention resource allocation. *Accid Anal Prev*. 2004;36(3):341-350.

107. Wickens CD. Multiple resources and performance prediction. *Theor Issues Ergon Sci*. 2002;3(2):159-177.

108. Jones DG, Endsley MR. Sources of situation awareness errors in aviation. *Aviat Space Environ Med*. 1996;67(6):507-512.

109. Gugerty LJ. *Situation Awareness during Driving: Explicit and Implicit Knowledge in Dynamic Spatial Memory.* Vol 3. US: American Psychological Association; 1997:42-66.

110. Catchpole KR, Giddings AEB, Hirst G, Dale T, Peek GJ, de Leval MR. A method for measuring threats and errors in surgery. *Cogn Technol Work*. 2008;10(4):295-304.

111. Endsley M, Bolstad C. Individual Differences in Pilot Situation Awareness. *Int J Aviat Psychol*. 1994;4(3):241-264.

112. Horswill MS, McKenna F. Drivers' hazard perception ability: Situation awareness on the road. In: Banbury S, Tremblay S, eds. *A Cognitive Approach to SItuation Awareness*. Aldershot, England: Ashgate Publishing; 2004:155-175.

113. Baker DP, Day R, Salas E. Teamwork as an Essential Component of High-Reliability Organizations. *Health Serv Res*. 2006;41(4):1576-1598.

114. Weaver SJ, Lyons R, DiazGranados D, et al. The anatomy of health care team training and the state of practice: a critical review. *Acad Med*. 2010;85(11):1746-1760.

115. Salas E, Dickinson T, Converse S, Tannenbaum S. Toward an understanding of team performance and training. In: Swezey R, Salas E, eds. *Teams: Their Training and Performance*. Norwood, NJ; 1992:3-29.

116. Baker DP, Gustafson S, Beaubien J, Salas E, Barach P. *Medical Teamwork and Patient Safety: The Evidence-Based Relation.* Rockville MD; 2005.

117. She M, Li Z. Team Situation Awareness: A Review of Definitions and Conceptual Models. In: Springer, Cham; 2017:406-415.

118. Prince C, Salas E. Team situation awareness, errors, and crew resource management: research integration for training guidance. In: Endsley M, Garland D, eds. *Situation Awareness Analysis and Measurement.* Mahwah, New Jersey: CRC Press; 2000:325-347.

119. Jentsch F, Barnett J, Bowers C. Loss of Aircrew Situation Awareness: A Cross-Validation. *Proc Hum Factors Ergon Soc Annu Meet*. 1997;41(2):1379-1379.

120. Hoeft R, Jentsch F, Smith-Jentsch K, Bowers C. Exploring the role of shared mental models for implicit coordination in teams. In: *Proceedings of the Human Factors and Ergonomics Society 49*

Th Annual Meeting. ; 2005:1863-1867.

121. Endsley MR, Jones DG. Designing to Support SA for Multiple and Distributed Operators. In: Endsley MR, Jones DG, eds. *Designing for Situation Awareness*. Second. Boca Raton, FL, US: CRC Press; 2012:193-218.

122. Shinar D. Traffic safety and inidividual differences in drivers' attention and information processing capacity. *Alcohol, Drugs Driv*. 1993;9:219-237.

123. Ranney TA. Models of driving behavior: a review of their evolution. *Accid Anal Prev*. 1994;26(6):733-750.

124. Chopra V, Bovill JG, Spierdijk J, Koornneef F. Reported Significant Observations During Anaesthesia: A Prospecitive Analysis Over an 18-month Period. *Br J Anaesth*. 1992;68:13-17.

125. Myers JA, Powell DMC, Aldington S, et al. The impact of fatigue on the non-technical skills performance of critical care air ambulance clinicians. *Acta Anaesthesiol Scand*. 2017;61(10):1305-1313.

126. Bhananker SM, Posner KL, Cheney FW, Caplan RA, Lee LA, Domino KB. Injury and liability associated with monitored anesthesia care: a closed claims analysis. *Anesthesiology*. 2006;104(2):228-234.

127. Thimbleby H, Lewis A, Williams J. Making healthcare safer by understanding, designing and buying better IT. *Clin Med*. 2015;15(3):258-262.

128. Hörmann H-J, Soll H, Dudfield H, Banbury S. ESSAI -Training of Situation Awareness and Threat Management Techniques - Results of an Evaluation Study. In: *12th International Symposium on Aviation Psychology*. Dayton OH; 2003:14-17.

129. McKenna F, Crick J. *Hazard Perception in Drivers: A Methodology for Testing and Training*.; 1994.

130. Grayson GB, Sexton BF. The development of hazard perception testing Prepared for Road Safety Division, Department for Transport. https://trl.co.uk/sites/default/files/TRL558.pdf. Accessed March 9, 2019.

131. Mercier A, Kerhuel N, Stalnikiewitz B, et al. Measuring situation awareness in emergency settings: A systematic review of tools and outcomes. *Open Access Emerg Med*. 2010;2(1):7-16.

132. Taylor RM. Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In: *AGARD Conference Proceedings No 478, Situational Awareness in*

Aerospace Operations. Neuilly sur Seine: NATO-AGARD; 1990:3/1-3/17.

133. Endsley M. Direct Measurement of Situation Awareness: Validity and Use of SAGAT. In: Endsley M, Garland D, eds. *Situation Awareness Analysis and Measurement.* Mahwah, New Jersey: CRC Press; 2000:147-173.

134. Vidulich MA, Hughes ER. Testing a subjective metric of situation awareness. In: *Proceedings of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA, US: Human Factors Society; 1991:1307-1311.

135. Bell HH, Waag WL. Using observer ratings to assess situational awareness in in tactical air environments. In: Garland DJ, Endsley MR, eds. *Experimental Analysis and Measurement of Situation Awareness*. Daytona Beach: Embry-Riddle Aeronautical University Press; 1995:93-99.

136. Hogan MP, Pace DE, Hapgood J, et al. Use of human patient simulation and the Situation Awareness Global Assessment Technique in practical trauma skills assessment. *J Trauma*. 2006;61(5):1047-1052.

137. Levin S, Sauer L, Kelen G, et al. Situation awareness in emergency medicine. *IIE Trans Healthc Syst Eng*. 2012;2(2):172-180.

138. Messick S. Validity. In: Linn R, ed. *Educational Measurement*. 3rd ed. Macmillan; 1988.

139. Cook DA, Beckman TJ. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *Am J Med*. 2006;119:166.e7-166.e16.

140. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Heal Sci Educ.* 2014;19(2):233-250.

141. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38(9):1006-1012.

142. Bruce T. Understanding Reliability and Coefficient alpha, Really In: Score Reliability. 2011.

143. Endsley M, Jones D. Determining SA requirements. In: *Designing for Situation Awareness*. Second. Boca Raton, FL, US: CRC Press; 2012:63-78.

144. Salmon P, Stanton N, Walker G, Green D. Situation awareness measurement: A review of applicability for C4i environments. *Appl Ergon*. 2006;37(2):225-238.

145. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology*. 1998;89(1):8-18.

146. Helmreich RL, Wilhelm JA. Outcomes of crew resource management training. *Int J Aviat Psychol*. 1991;1(4):287-300.

147. Endsley MR. Direct measurement of situation awareness: Validity and use of SAGAT. *Situat Aware Anal Meas*. 2000:147-173.

148. Zhang Y, Drews FA, Westenskow DR, et al. Effects of Integrated Graphical Displays on Situation Awareness in Anaesthesiology. *Cogn Technol Work*. 2002;4:82-90.

149. Levin S, Sauer L, Kelen G, Kirsch T, Pham J, Desai S. Situation awareness in emergency medicine. *IIE Trans Healthc Syst Eng*. 2012;2(2):172-180.

150. Morgan P, Tregunno D, Brydges R, et al. Using a situational awareness global assessment technique for interprofessional obstetrical team training with high fidelity simulation. *J Interprof Care*. 2015;29(1):13-19.

151. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of Physician Self-assessment Compared With Observed Measures of Competence. *JAMA*. 2006;296(9):1094-1102.

152. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth*. 2003;90(5):580-588.

153. Yule S, Flin R, Maran N, et al. Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System. *World J Surg*. 2008;32(4):548-556.

154. Cooper S, Cant R, Porter J, et al. Rating medical emergency teamwork performance: Development of the Team Emergency Assessment Measure (TEAM). *Resuscitation*. 2010;81(4):446-452.

155. Endsley MR, Jones DG. *Designing for Situation Awareness : An Approach to User-Centered Design*. CRC Press; 2012.

156. Mcgaghie WC, Issenberg SB, Barsuk JH, Wayne DB. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ*. 2014;48(4):375-385.

157. Gaba DM. Improving Anesthesiologists' Performance by SImulating Reality. *Anesthesiology*.

1992;76(4):491-494.

158. Gaba DM. The future vision of simulation in health care. *Qual Saf Heal Care*. 2004;13:2-10.

159. Graber MA, Wyatt C, Kasparek L, Xu Y. Does Simulator Training for Medical Students Change Patient Opinions and Attitudes toward Medical Student Procedures in the Emergency Department? *Acad Emerg Med*. 2005;12(7):635-639.

160. Salas E, Rosen A. M. Building high reliability teams: progress and some reflections on teamwork training. *BMJ Qual Saf*. 2013;22(5):369-373.

161. Grand A. J, Pearce M, Rench A. T, et al. Going DEEP: guidelines for building simulation-based team assessments. *BMJ Qual Saf*. 2013;22(5):436-448.

162. Neily J, Mills PD, Young-Xu Y, et al. Association Between Implementation of a Medical Team Training Program and Surgical Mortality. *JAMA*. 2010;304(15):1693.

163. Armour Forse R, Bramble JD, McQuillan R. Team training can improve operating room performance. *Surgery*. 2011;150(4):771-778.

164. Deering S, Rosen MA, Ludi V, et al. On the front lines of patient safety: Implementation and evaluation of team training in Iraq. *Jt Comm J Qual Patient Saf*. 2011;37(8):350-356.

165. Riley W, Davis S, Miller K, Hansen H, Sainfort F, Sweet R. Didactic and simulation nontechnical skills team training to improve perinatal patient outcomes in a community hospital. *Jt Comm J Qual Patient Saf*. 2011;37(8):357-364.

166. Young-Xu Y, Neily J, Mills PD, et al. Association Between Implementation of a Medical Team Training Program and Surgical Morbidity. *Arch Surg*. 2011;146(12):1368.

167. Walker GH, Stanton NA, Kazi TA, Salmon PM, Jenkins DP. Does advanced driver training improve situational awareness? *Appl Ergon*. 2009;40(4):678-687.

168. Endsley M, Jones D. SA Oriented Training. In: Endsley M, Jones D, eds. *Designing for Situation Awareness*. Second. CRC Press; 2011:235-258.

169. Mishra A, Catchpole K, Dale T, McCulloch P. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surg Endosc*. 2008;22(1):68-73.

170. Singh H, Davis Giardina T, Petersen LA, et al. Exploring situational awareness in diagnostic errors in primary care. *BMJ Qual Saf*. 2012;1:30-38.

171. Sitterding MC, Ebright P, Broome M, Patterson ES, Wuchner S. Situation Awareness and Interruption Handling During Medication Administration. *West J Nurs Res*. 2014;36(7):891-916.

172. Endsley MR. Measurement of situation awareness in dynamic systems. *Spec Issue Situat Aware*. 1995;37(1):65-84.

173. Helmreich RL. On error management : lessons from aviation. *BMJ*. 2002;320(18):781-785.

174. Woloshynowych M, Rogers S, Taylor-Adams S, Vincent C. The investigation and analysis of critical incidents and adverse events in healthcare. *Health Technol Assess (Rockv)*. 2005;9(19).

175. Vincent C. Understanding how things go wrong. In: *Patient Safety*. Wiley-Blackwell; 2010:141-167.

176. Amalberti R, Benhamou D, Auroy Y, Degos L. Adverse events in medicine: Easy to count, complicated to understand, and complex to prevent. *J Biomed Inform*. 2011;44(3):390-394.

177. NHS England. Serious Incident Framework Supporting learning to prevent recurrence. 2015:12-16.

178. Mason J. *Qualitative Researching*. Second. London: Sage Publications; 2002.

179. Ritchie J, Lewis J, McNaughton Nicholls C, Ormston R. *Qualitative Research Practice : A Guide for Social Science Students and Researchers*. London: Sage Publications Inc.; 2014.

180. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77-101.

181. World Health Organisation. *Conceptual Framework for the International Classification for Patient Safety*.; 2009. http://www.who.int/patientsafety/taxonomy/ICPS_Statement_of_Purpose.pdf. Accessed March 9, 2019.

182. Chang A, Schyve PM, Croteau RJ, O'Leary DS, Loeb JM. The JCAHO patient safety event taxonomy: a standardized terminology and classification schema for near misses and adverse events. *Int J Qual Heal Care*. 2005;17(2):95-105.

183. National Health Service. National Reporting and Learning System. https://report.nrls.nhs.uk/nrlsreporting/. Accessed March 9, 2019.

184. Taylor-Adams S, Vincent C. Systems analysis of clinical incidents: the London protocol. *Clin Risk*. 2004;10(6):211-220.

185. Rasmussen J. *Information Processing and Human-Machine Interaction : An Approach to Cognitive*

Engineering. North-Holland; 1986.

186. Sarter NB, Alexander HM. Error Types and Related Error Detection Mechanisms in the Aviation Domain: An Analysis of Aviation Safety Reporting System Incident Reports. *Int J Aviat Psychol*. 2000;10(2):189-206.

187. Tallentire VR, Smith SE, Skinner J, Cameron HS. Exploring Error in Team-Based Acute Care Scenarios: An Observational Study From the United Kingdom. *Acad Med Acad Med*. 2012;8787(6):792-798.

188. Schulz CM, Endsley MR, Kochs EF, Gelb AW, Wagner KJ. Situation Awareness in Anesthesia. *Anesthesiology*. 2013;118(3):729-742.

189. NCEPOD. *Themes and Recommendations Common to All Hospital Specialities*.; 2018. www.hqip.org.uk/national-programmes. Accessed March 9, 2019.

190. Han L, Sutton M, Clough S, Warner R, Doran T. Impact of out-of-hours admission on patient mortality: longitudinal analysis in a tertiary acute hospital. *BMJ Qual Saf*. 2018;27:445-454.

191. Freemantle N, Richardson M, Wood J, et al. Weekend hospitalization and additional risk of death: An analysis of inpatient data. *J R Soc Med*. 2012;105(2):74-84.

192. Freemantle N, Ray D, McNulty D, et al. Increased mortality associated with weekend hospital admission: a case for expanded seven day services? *BMJ*. 2015;351:h4596.

193. Aylin P, Alexandrescu R, Jen MH, Mayer EK, Bottle A. Day of week of procedure and 30 day mortality for elective surgery: retrospective analysis of hospital episode statistics. *BMJ*. 2013;346:f2424.

194. Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. *Med J Aust*. 1995;163(9):458-471.

195. Leape LL, Brennan TA, Laird N, et al. The Nature of Adverse Events in Hospitalized Patients. *N Engl J Med*. 1991;324(6):377-384.

196. Bellomo R, Goldsmith D, Russell S, Uchino S. Postoperative Serious Adverse Events in a Teaching Hospital: A Prospective Study. *Med J Aust*. 2002;176(5):216-218.

197. Rothschild JM, Bates DW, Leape LL. Preventable Medical Injuries in Older Patients. *Arch Intern Med*. 2000;160(18):2717.

198. Reason J. Safety in the operating theatre-Part 2: Human error and organisational failure. *Curr Anaesth Crit Care*. 1995;6:121-126.

199. Schaefer HG, Helmreich RL, Scheidegger D. Safety in the operating theatre-Part 1: Interpersonal relationships and team performance. *Curr Anaesth Crit Care*. 1995;6:48-53.

200. Andersen PO, Maaløe R, Andersen HB. Critical incidents related to cardiac arrests reported to the Danish Patient Safety Database. *Resuscitation*. 2010;81(3):312-316.

201. Undre S, Sevdalis N, Healey AN, Darzi SA, Vincent CA. Teamwork in the operating theatre: cohesion or confusion? *J Eval Clin Pract*. 2006;12(2):182-189.

202. Reader T, Flin R, Lauche K, Cuthbertson BH. Non-technical skills in the intensive care unit. *Br J Anaesth Br J Anaesth*. 2006;96(96):551-559.

203. Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiol Scand*. 2009;53(2):143-151.

204. Lingard L. Communication failures in the operating room: an observational classification of recurrent types and effects. *Qual Saf Heal Care*. 2004;13(5):330-334.

205. Tom W Reader, Rhona Flin BHC. Communication skills and error in the intensive care unit. *Curr Opin Crit Care*. 2007;13(6):732-736.

206. Nemeth CP. The Context for Improving Healthcare Team Communication. In: Nemeth CP, ed. *Improving Healthcare Team Communication*. Ashgate Publishing; 2008:1-7.

207. Wears RL, Woloshynowych M, Brown R, Vincent CA. Reflective analysis of safety research in the hospital accident and emergency departments. *Appl Ergon*. 2010;41(5):695-700.

208. Wright MC, Endsley MR. Building shared situation awareness in healthcare setting. In: Nemeth CP, ed. *Improvng Healthcare Team Communication*. Ashgate Publishing; 2008:98-114.

209. Sexton JB, Thomas EJ, Helmreich RL. Error, stress, and teamwork in medicine and aviation: cross sectional surveys. *BMJ*. 2000;320:745-749.

210. Bion J, Heffner J. Challenges in the care of the acutely ill. *Lancet*. 2004;363(9413):970-977.

211. Weigl M, Mü Ller A, Vincent C, Angerer P, Sevdalis N. The association of workflow interruptions and hospital doctors' workload: a prospective observational study. *BMJ Qual Saf*. 2011;21:399-407.

212. Gawande AA, Thomas EJ, Zinner MJ, Brennan TA. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*. 1999;126(1):66-75.

213. Shappell S, Wiegmann D. Applying reason: The human factors analysis and classification system (HFACS). *Hum Factors Aerosp Saf*. 2001;1(1):59-86.

214. van der Schaaf TW. Development of a near miss management system at a chemical process plant. In: van der Schaaf T, Lucas D, Hale A, eds. *Near Miss Reporting as a Safety Tool*. Butterworth-Heinemann; 1991:57-63.

215. Svenson O. The Accident Evolution and Barrier Function (AEB) Model Applied to Incident Analysis in the Processing Industries. *Risk Anal*. 1991;11(3):499-507.

216. Dornan T, Ashcroft D, Heathfield H, et al. *An in Depth Investigation into Causes of Prescribing Errors by Foundation Trainees in Relation to Their Medical Education*. https://www.gmc-uk.org/FINAL_Report_prevalence_and_causes_of_prescribing_errors.pdf_28935150.pdf. Accessed March 9, 2019.

217. Stiegler MP, Neelankavil JP, Canales C, Dhillon A. Cognitive errors detected in anaesthesiology: a literature review and pilot study. *Br J Anaesth*. 2012;108(2):229-235.

218. Tokuda Y, Kishida N, Konishi R, Koizumi S, Koizumi S. Cognitive error as the most frequent contributory factor in cases of medical injury: A study on verdict's judgment among closed claims in Japan. *J Hosp Med*. 2011;6(3):109-114.

219. Reason J. Performance levels and error types. In: *Human Error*. Cambridge University Press; 1990:53-96.

220. Reason J. The Detection of Errors. In: *Human Error*. Cambridge University Press; 1990:148-172.

221. Rothschild JM, Landrigan CP, Cronin JW, et al. The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care*. *Crit Care Med*. 2005;33(8):1694-1700.

222. Tallentire VR, Smith SE, Skinner J, Cameron HS. Exploring patterns of error in acute care using framework analysis. *BMC Med Educ.* 2015;15(1):3.

223. Reason J. Latent errors and systems disasters. In: *Human Error*. Cambridge University Press; 1990:173-216.

224. Dörner D. On the Difficulties People have in Dealing with Complexity. *Simul Games*.

1980;11(1):87-106.

225. Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K. Nurse-Staffing Levels and the Quality of Care in Hospitals. *N Engl J Med*. 2002;346(22):1715-1722.

226. Dixon-Woods M, Baker R, Charles K, et al. Culture and behaviour in the English National Health Service: overview of lessons from a large multimethod study. *BMJ Qual Saf*. 2014;23(2):106-115.

227. Parasuraman R, Riley V. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum Factors J Hum Factors Ergon Soc*. 1997;39(2):230-253.

228. Vincent C. *Patient Safety*. 2nd ed. Wiley-Blackwell Publishing Ltd.; 2010.

229. Ulrich RS, Zimring C, Zhu X, et al. A Review of the Research Literature on Evidence-Based Healthcare Design. *HERD Heal Environ Res Des J*. 2008;1(3):61-125.

230. Haynes AB, Weiser TG, Berry WR, et al. A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population. *N Engl J Med*. 2009;360(5):491-499.

231. NHS Improvement. National Safety Standards for Invasive Procedures. https://improvement.nhs.uk/documents/923/natssips-safety-standards.pdf. Published 2015. Accessed March 9, 2019

232. Whyte S, Lorelei AE, Ae L, et al. Paradoxical effects of interprofessional briefings on OR team performance. *Cogn Technol Work*. 2008;10:287-294.

233. Hannam JA, Glass L, Kwon J, et al. A prospective, observational study of the effects of implementation strategy on compliance with a surgical safety checklist. *BMJ Qual Saf*. 2013;22(11):940-947.

234. Russ S, Rout S, Caris J, et al. Measuring Variation in Use of the WHO Surgical Safety Checklist in the Operating Room: A Multicenter Prospective Cross-Sectional Study. *J Am Coll Surg*. 2015;220(1):1-11.e4.

235. Howell A-M, Burns EM, Bouras G, Donaldson LJ, Athanasiou T, Darzi A. Can Patient Safety Incident Reports Be Used to Compare Hospital Safety? Results from a Quantitative Analysis of the English National Reporting and Learning System Data. *PLoS One*. 2015;10(12):e0144107.

236. Vincent CA. Reporting and learning systems. In: *Patient Safety*. 2nd ed. Wiley-Blackwell; 2010:75-95.

237. Anderson JE, Kodate N, Walters R, Dodds A. Can incident reporting improve safety? Healthcare practitioners' views of the effectiveness of incident reporting. *Int J Qual Heal Care*. 2013;25(2):141-150.

238. Stavropoulou C, Doherty C, Tosey P. How Effective Are Incident-Reporting Systems for Improving Patient Safety? A Systematic Literature Review. *Milbank Q*. 2015;93(4):826-866.

239. Macrae C. The problem with incident reporting. *BMJ Qual Saf*. 2016;25(2):71-75.

240. Leaver M, Reader TW. Human Factors in Financial Trading: An Analysis of Trading Incidents. *Hum Factors*. 2016;58(6):814-832.

241. Hewitt T, Chreim S, Forster A. Incident reporting systems: a comparative study of two hospital divisions. *Arch Public Health*. 2016;74:34.

242. House of Commons Public Administration Select Committee. Investigating clinical incidents in the NHS. 2015;(HC886). https://publications.parliament.uk/pa/cm201415/cmselect/cmpubadm/886/886.pdf. Accessed March 9, 2019

243. Michel P, Brami J, Chanelière M, et al. Patient safety incidents are common in primary care: A national prospective active incident reporting survey. *PLoS One*. 2017;12(2):e0165455.

244. Thomas EJ, Sexton JB, Helmreich RL. Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation. *Qual Saf Heal Care*. 2004;13:57-64.

245. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-Enhanced Simulation to Assess Health Professionals: A Systematic Review of Validity Evidence, Research Methods, and Reporting Quality. *Acad Med*. 2013;88(6):872-883.

246. Hawker S, Payne S, Kerr C, Hardey M, Powell J. Appraising the evidence: reviewing disparate data systematically. *Qual Health Res*. 2002;12(9):1284-1299.

247. Gale TCE, Roberts MJ, Sice PJ, et al. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth*. 2010;105(5):603-609.

248. Undre S, Healey AN, Darzi A, Vincent CA. Observational assessment of surgical teamwork: A feasibility study. *World J Surg*. 2006;30(10):1774-1783.

249. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool

for measuring teamwork behaviour in the operating theatre. *Qual Saf Heal Care*. 2009;18(2):104-108.

250. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ*. 2006;40(11):1098-1104.

251. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. 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. *Crit Care Med*. 2006;34(8):2167-2174.

252. Malec JF, Torsher LC, Dunn WF, et al. The Mayo High Performance Teamwork Scale : Reliability and Validity for Evaluating Key Crew Resource Management Skills. *Simul Healthc*. 2007;2(1):4-10.

253. Weller J, Frengley R, Torrie J, et al. Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. *BMJ Qual Saf*. 2011;20(January):216-222.

254. Blackhall LJ, Erickson J, Brashers V, Owen J, Thomas S. Development and Validation of a Collaborative Behaviors Objective Assessment Tool for End-of-Life Communication. *J Palliat Med*. 2014;17(1):68-74.

255. Dedy NJ, Szasz P, Louridas M, Bonrath EM, Husslein H, Grantcharov TP. Objective structured assessment of nontechnical skills: Reliability of a global rating scale for the in-training assessment in the operating room. *Surg (United States)*. 2015;157(6):1002-1013.

256. Tregunno D, Pittini R, Haley M, Morgan PJ, Tregunno D. Development and usability of a behavioural marking system for performance assessment of obstetrical teams. *Qual Saf Heal Care*. 2009;18:393-396.

257. Raison N, Wood T, Brunckhorst O, et al. Development and validation of a tool for non-technical skills evaluation in robotic surgery-the ICARS system. *Surg Endosc*. 2017;31(12):5403-5410.

258. Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: review of published studies. *BJA Br J Anaesth*. 2001;86(3):445-450.

259. Kardong-Edgren S, Adamson KA, Fitzgerald C. A Review of Currently Published Evaluation Instruments for Human Patient Simulation. *Clin Simul Nurs*. 2010;6(1):e25-e35.

260. American Educational Research Association., American Psychological Association., National Council on Measurement in Education., Joint Committee on Standards for Educational and

Psychological Testing (U.S.). *Standards for Educational and Psychological Testing*. American Educational Research Association; 1999.

261. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326(7379):41-44.

262. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96-106.

263. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3):309-317.

264. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63:737-745.

265. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul*. 2016;1(1):31.

266. Salas E, Almeida SA, Salisbury M, et al. What are the critical success factors for team training in health care? *Jt Comm J Qual Patient Saf*. 2009;35(8):398-405.

267. Rutherford JS, Flin R, Irwin A, McFadyen AK. Evaluation of the prototype Anaesthetic Non-technical Skills for Anaesthetic Practitioners (ANTS-AP) system: a behavioural rating system to assess the non-technical skills used by staff assisting the anaesthetist. *Anaesthesia*. 2015;70(8):907-914.

268. Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent CA. Reliability of a revised NOTECHS scale for use in surgical teams. *Am J Surg*. 2008;196(2):184-190.

269. Undre S, Healey AN, Darzi A, Vincent CA. Observational Assessment of Surgical Teamwork: A Feasibility Study. *World J Surg*. 2006;30(10):1774-1783.

270. Flowerdew L, Brown R, Vincent C, Woloshynowych M. Development and validation of a tool to assess emergency physicians' nontechnical skills. *Ann Emerg Med*. 2012;59(5):376-385.e4.

271. Amaya Arias AC, Barajas R, Eslava-Schmalbach JH, et al. Translation, cultural adaptation and content re-validation of the observational teamwork assessment for surgery tool. *Int J Surg*. 2014;12(12):1390-1402.

272. Amiel I, Simon D, Merin O, Ziv A. Mobile in Situ Simulation as a Tool for Evaluation and

Improvement of Trauma Treatment in the Emergency Department. *J Surg Educ*. 2016;73(1):121-128.

273. Spanager L, Lyk-Jensen HT, Dieckmann P, Wettergren A, Rosenberg J, Ostergaard D. Customization of a tool to assess Danish surgeons' non-technical skills in the operating room. *Dan Med J*. 2012;59(11):1-6.

274. Passauer-Baierl S, Hull L, Miskovic D, Russ S, Sevdalis N, Weigl M. Re-validating the observational teamwork assessment for surgery tool (OTAS-D): Cultural adaptation, refinement, and psychometric evaluation. *World J Surg*. 2014;38(2):305-313.

275. Maignan M, Koch F-XF-X, Chaix J, et al. Team Emergency Assessment Measure (TEAM) for the assessment of non-technical skills during resuscitation: Validation of the french version. *Resuscitation*. 2016;101:115-120.

276. Michinov E, Jamet E, Dodeler V, Haegelen C, Jannin P. Assessing neurosurgical non-technical skills: An exploratory study of a new behavioural marker system. *J Eval Clin Pract*. 2014;20(5):582-588.

277. Kemper PF, van Noord I, de Bruijne M, et al. Development and reliability of the explicit professional oral communication observation tool to quantify the use of non-technical skills in healthcare. *BMJ Qual Saf*. 2013;22(7):586-595.

278. Brashers V, Erickson JM, Blackhall L, Owen JA, Thomas SM, Conaway MR. Measuring the impact of clinically relevant interprofessional education on undergraduate medical and nursing student competencies: A longitudinal mixed methods approach. *J Interprof Care*. 2016;30(4):448-457.

279. Anderson JM, Murphy AA, Boyle KB, Yaeger KA, Halamek LP. Simulating extracorporeal membrane oxygenation emergencies to improve human performance. Part II: assessment of technical and behavioral skills. *Simul Healthc*. 2006;1(4):228-232.

280. Parker-Raley J, Cerroni A, Mottet TP, et al. Investigating pediatric trauma team communication effectiveness phase two: Achieving inter-rater reliability for the Assessment of Pediatric Resuscitation Communication Team Assessment. *J Commun Healthc*. 2013;6(3):145-157.

281. Weller JM, Jolly B, Robinson B. Generalisability of behavioural skills in simulated anaesthetic emergencies. *Anaesth Intensive Care*. 2008;36(2):185-189.

282. Frankel A, Gardner R, Maynard L, et al. Using the Communication and Teamwork Skills (CATS) assessment to measure health care team performance. *Jt Comm J Qual Patient Saf*.

2007;33(9):549-558.

283. Grant EC, Grant VJ, Bhanji F, Duff JP, Cheng A, Lockyer JM. The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. *Resuscitation*. 2012;83(7):887-893.

284. O'Leary KJ, Boudreau YN, Creden AJ, Slade ME, Williams M V. Assessment of teamwork during structured interdisciplinary rounds on medical units. *J Hosp Med*. 2012;7(9):679-683.

285. Huang LC, Conley D, Lipsitz S, et al. The surgical safety checklist and teamwork coaching tools: A study of inter-rater reliability. *BMJ Qual Saf*. 2014;23(8):639-650.

286. Pugh D, Hamstra SJ, Wood TJ, et al. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Adv Heal Sci Educ*. 2014;20(1):85-100.

287. Franc JM, Verde M, Gallardo AR, Carenzo L, Ingrassia PL. An Italian version of the Ottawa Crisis Resource Management Global Rating Scale: a reliable and valid tool for assessment of simulation performance. *Intern Emerg Med*. 2016;12(5):651-656.

288. Holly D, Swanson V, Cachia P, Beasant B, Laird C. Development of a behaviour rating system for rural/remote pre-hospital settings. *Appl Ergon*. 2017;58:405-413.

289. Ottestad E, Boulet JR, Lighthall GK. Evaluating the management of septic shock using patient simulation. *Crit Care Med*. 2007;35(3):769-775.

290. Carlson J, Min E, Bridges D. The impact of leadership and team behavior on standard of care delivered during human patient simulation: a pilot study for undergraduate medical students. *Teach Learn Med*. 2009;21(1):24-32.

291. Hamilton N, Freeman B, Woodhouse J, Ridley C, Murray D, Klingensmith ME. Team Behavior During Trauma Resuscitation: A Simulation-Based Performance Assessment. *J Grad Med Educ*. 2009;(December):253-259.

292. Hamilton NA, Kieninger AN, Woodhouse J, Freeman BD, Murray D, Klingensmith ME. Video review using a reliable evaluation metric improves team function in high-fidelity simulated trauma resuscitation. *J Surg Educ*. 2012;69(3):428-431.

293. Paige JT, Garbee DD, Kozmenko V, et al. Getting a Head Start: High-Fidelity, Simulation-Based Operating Room Team Training of Interprofessional Students. *J Am Coll Surg*. 2014;218(1):140-149.

294. DeMoor S, Abdel-Rehim S, Olmsted R, Myers JG, Parker-Raley J. Evaluating trauma team performance in a Level I trauma center. *J Trauma Acute Care Surg*. 2017;83(1):159-164.

295. Wright MC, Phillips-Bute BG, Petrusa ER, Griffin KL, Hobbs GW, Taekman JM. Assessing teamwork in medical education and practice: Relating behavioural teamwork ratings and clinical performance. *Med Teach*. 2009;31(1):30-38.

296. Reid J, Stone K, Brown J, et al. The Simulation Team Assessment Tool (STAT): Development, reliability and validation. *Resuscitation*. 2012;83(7):879-886.

297. Pascual JL, Holena DN, Vella MA, et al. Short Simulation Training Improves Objective Skills in Established Advanced Practitioners Managing Emergencies on the Ward and Surgical Intensive Care Unit. *J Trauma Inj Infect Crit Care*. 2011;71(2):330-338.

298. Von Wyl T, Zuercher M, Amsler F, Walter B, Ummenhofer W. Technical and non-technical skills can be reliably assessed during paramedic simulation training. *Acta Anaesthesiol Scand*. 2008;53(1):121-127.

299. Daniels K, Lipman S, Harney K, Arafeh J, Druzin M. Use of Simulation Based Team Training for Obstetric Crises in Resident Education. *Simul Healthc J Soc Simul Healthc*. 2008;3(3):154-160.

300. Arain NA, Hogg DC, Gala RB, et al. Construct and face validity of the American College of Surgeons/Association of Program Directors in Surgery laparoscopic troubleshooting team training exercise. *Am J Surg*. 2012;203(1):54-62.

301. Weller JM, Bloch M, Young S, et al. Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. *Br J Anaesth*. 2003;90(1):43-47.

302. Weller JM, Robinson BJ, Jolly B, et al. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anaesthesia*. 2005;60(3):245-250.

303. Kane MT. An argument-based approach to validity. *Psychol Bull*. 1992;112(3):527-535.

304. Gale TCE, Roberts MJ, Sice PJ, et al. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth*. 2010;105(5):603-609.

305. Morgan PJ, Tregunno D, Pittini R, et al. Determination of the psychometric properties of a behavioural marking system for obstetrical team training using high-fidelity simulation. *BMJ Qual Saf*. 2012;21(1):78-82.

306. Lyk-Jensen HT, Dieckmann P, Konge L, et al. Using a Structured Assessment Tool to Evaluate Nontechnical Skills of Nurse Anesthetists. *AANA J*. 2016;84(2):122-127.

307. Yule S, Flin R, Maran N, et al. Debriefing surgeons on non-technical skills (NOTSS). *Cogn Technol Work*. 2008;10(4):265-274.

308. Balki M, Hoppe D, Monks D, et al. The PETRA (Perinatal Emergency Team Response Assessment) Scale: A High-Fidelity Simulation Validation Study. *J Obstet Gynaecol Canada*. 2017;39(7):523-533.e12.

309. Mitchell L, Flin R, Yule S, Mitchell J, Coutts K, Youngson G. Evaluation of the Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS) system. *Int J Nurs Stud*. 2011;49(2):201-211.

310. Olupeliyawa AM, O'Sullivan AJ, Hughes C, Balasooriya CD. The Teamwork Mini-Clinical Evaluation Exercise (T-MEX): A workplace-based assessment focusing on collaborative competencies in health care. *Acad Med*. 2014;89(2):359-365.

311. Thistlethwaite J, Dallest K, Moran M, et al. Introducing the individual Teamwork Observation and Feedback Tool (iTOFT): Development and description of a new interprofessional teamwork measure. *J Interprof Care*. 2016;30(4):526-528.

312. Jepsen RMHGMHG, Dieckmann P, Spanager L, et al. Evaluating structured assessment of anaesthesiologists' non-technical skills. *Acta Anaesthesiol Scand*. 2016;60(6):756-766.

313. Russ S, Hull L, Rout S, Vincent C, Darzi A, Sevdalis N. Observational teamwork assessment for surgery: Feasibility of clinical and nonclinical assessor calibration with short-term training. *Ann Surg*. 2012;255(4):804-809.

314. Bracco F, Masini M, De Tonetti G, et al. Adaptation of non-technical skills behavioural markers for delivery room simulation. *BMC Pregnancy Childbirth*. 2017;17(1):1-7.

315. Guise J-M, Deering SH, Kanki BG, Osterweil P, Li H, Mori M. Validation of a Tool to Measure and Promote Clinical Teamwork. *Simul Healthc*. 2008;3(4):217-223.

316. Cooper SJ, Cant RP. Measuring non-technical skills of medical emergency teams: An update on the validity and reliability of the team emergency assessment measure (TEAM). *Resuscitation*. 2014;85(1):31-33.

317. Cooper S, Cant R, Connell C, et al. Measuring teamwork performance: Validity testing of the Team Emergency Assessment Measure (TEAM) with clinical resuscitation teams. *Resuscitation*.

2016;101:97-101.

318. Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. *Qual Saf Heal Care*. 2004;13(SUPPL. 1):i33-i40.

319. Undre S, Sevdalis N, Healey AN, Darzi A, Vincent CA. Observational Teamwork Assessment for Surgery (OTAS): Refinement and Application in Urological Surgery. *World J Surg*. 2007;31(7):1373-1381.

320. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: Construct validation with expert versus novice raters. *Ann Surg*. 2009;249(6):1047-1051.

321. Morgan LJ, Pickering SP, Collins G, et al. Oxford NOTECHS II: A modified theatre team non-technical skills scoring system. *PLoS One*. 2014;9(3):e90320.

322. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. *Cogn Technol Work*. 2004;6(3):165-171.

323. Patey R, Flin R, Fletcher G, Maran N, Glavin R. Developing a Taxonomy of Anesthetists' Nontechnical Skills (ANTS). *Adv Patient Saf from Res to Implement*. 2005;4:325-336.

324. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. 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. *Crit Care Med*. 2006;34(8):2167-2174.

325. Kim J, Neilipovitz D, Cardinal P, Chiu M. A Comparison of Global Rating Scale and Checklist Scores in the Validation of an Evaluation Tool to Assess Performance in the Resuscitation of Critically Ill Patients During Simulated Emergencies (Abbreviated as "CRM Simulator Study IB"). *Simul Healthc J Soc Simul Healthc*. 2009;4(1):6-16.

326. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery*. 2006;139(2):140-149.

327. Balki M, Hoppe D, Monks D, et al. Multidisciplinary Delphi Development of a Scale to Evaluate Team Function in Obstetric Emergencies: The PETRA Scale. *J Obstet Gynaecol Can*. 2017;39(6):434-442.e2.

328. Thomas EJ, Sexton JB, Lasky RE, Helmreich RL, Crandell DS, Tyson J. Teamwork and quality during neonatal care in the delivery room. *J Perinatol*. 2006;26(3):163-169.

329. Thomas EJ, Taggart B, Crandell S, et al. Teaching teamwork during the Neonatal Resuscitation Program: a randomized trial. *J Perinatol*. 2007;27(7):409-414.

330. Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF. Evaluating teamwork in a simulated obstetric environment. *Anesthesiology*. 2007;106(5):907-915.

331. Schraagen JM, Schouten T, Smit M, et al. Assessing and improving teamwork in cardiac surgery. *Qual Saf Health Care*. 2010;19(6):e29.

332. Schraagen JM, Schouten T, Smit M, et al. A prospective study of paediatric cardiac surgical microsystems: assessing the relationships between non-routine events, teamwork and patient outcomes. *BMJ Qual Saf*. 2011;20(7):599-603.

333. Mitchell L, Flin R, Yule S, Mitchell J, Coutts K, Youngson G. Development of a behavioural marker system for scrub practitioners' non-technical skills (SPLINTS system). *J Eval Clin Pract*. 2012;19(2):317-323.

334. Sutton G, Liao J, Jimmieson NL, Restubog SLD. Measuring Multidisciplinary Team Effectiveness in a Ward-Based Healthcare Setting: Development of the Team Functioning Assessment Tool. *J Healthc Qual Promot Excell Healthc*. 2011;33(3):10-24.

335. Sutton G, Liao J, Jimmieson NL, Restubog SLD. Measuring Ward-Based Multidisciplinary Healthcare Team Functioning: A Validation Study of the Team Functioning Assessment Tool (TFAT). *J Healthc Qual Promot Excell Healthc*. 2013;35(4):36-49.

336. Walker S, Brett S, McKay A, Lambden S, Vincent C, Sevdalis N. Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR): Development and validation. *Resuscitation*. 2011;82(7):835-844.

337. Devcich DA, Weller J, Mitchell SJ, et al. A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: Development and initial evaluation of the WHOBARS. *BMJ Qual Saf*. 2016;25(10):778-786.

338. Wright MC, Segall N, Hobbs G, Phillips-Bute B, Maynard L, Taekman JM. Standardized Assessment for Evaluation of Team Skills. *Simul Healthc J Soc Simul Healthc*. 2013;8(5):292-303.

339. Jepsen RMHG, Spanager L, Lyk-Jensen HT, Dieckmann P, Østergaard D. Customisation of an

instrument to assess anaesthesiologists' non-technical skills. *Int J Med Educ*. 2015;6:17-25.

340. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Ann Surg*. 2005;242(5):631-639.

341. Undre S, Koutantji M, Sevdalis N, et al. Multidisciplinary crisis simulations: The way forward for training surgical teams. *World J Surg*. 2007;31(9):1843-1853.

342. Koutantji M, McCulloch P, Undre S, et al. Is team training in briefings for surgical teams feasible in simulation? *Cogn Technol Work*. 2008;10(4):275-285.

343. Fernandez R, Kozlowski SWJ, Shapiro MJ, Salas E. Toward a Definition of Teamwork in Emergency Medicine. *Acad Emerg Med*. 2008;15(11):1104-1112.

344. Fernandez R, Pearce M, Grand J., et al. Evaluation of a Computer-Based Educational Intervention to Improve Medical Teamwork and Performance During Simulated Patient Resuscitations. *Crit Care Med*. 2013;41(11):2551-2562.

345. Crossingham G V., Sice PJA, Roberts MJ, Lam WH, Gale TCE. Development of workplace-based assessments of non-technical skills in anaesthesia. *Anaesthesia*. 2012;67(2):158-164.

346. Steinemann S, Berg B, DiTullio A, et al. Assessing teamwork in the trauma bay: Introduction of a modified "nOTECHS" scale for trauma. *Am J Surg*. 2012;203(1):69-75.

347. Lyk-Jensen HT, Jepsen RMHG, Spanager L, Dieckmann P, Østergaard D. Assessing nurse anaesthetists' non-technical skills in the operating room. *Acta Anaesthesiol Scand*. 2014;58(7):794-801.

348. Ahmed K, Anderson O, Jawad M, et al. Design and validation of the surgical ward round assessment tool: a quantitative observational study. *Am J Surg*. 2015;209:682-688.e2.

349. Hull L, Bicknell C, Patel K, et al. Content Validation and Evaluation of an Endovascular Teamwork Assessment Tool. *Eur J Vasc Endovasc Surg*. 2016;52(1):11-20.

350. Myers JA, Powell DMC, Psirides A, Hathaway K, Aldington S, Haney MF. Non-technical skills evaluation in the critical care air ambulance environment: introduction of an adapted rating instrument-an observational study. *Scand J Trauma Resusc Emerg Med*. 2016;24(1):24.

351. Flowerdew L, Gaunt A, Spedding J, et al. A multicentre observational study to evaluate a new tool to assess emergency physicians' non-technical skills. *Emerg Med J*. 2013;30(6):437-443.

352. Frengley RW, Weller JM, Torrie J, et al. The effect of a simulation-based training intervention on the performance of established critical care unit teams. *Crit Care Med*. 2011;39(12):2605-2611.

353. Weller J, Shulruf B, Torrie J, et al. Validation of a measurement tool for self-assessment of teamwork in intensive care. *Br J Anaesth*. 2013;111(3):460-467.

354. Spanager L, Beier-Holgersen R, Dieckmann P, Konge L, Rosenberg J, Oestergaard D. Reliable assessment of general surgeons' non-technical skills based on video-recordings of patient simulated scenarios. *Am J Surg*. 2013;206(5):810-817.

355. Spanager L, Konge L, Dieckmann P, Beier-Holgersen R, Rosenberg J, Oestergaard D. Assessing Trainee Surgeons' Nontechnical Skills: Five Cases are Sufficient for Reliable Assessments. *J Surg Educ*. 2015;72(1):16-22.

356. Mellanby E, Hume M, Glavin R, Skinner J, Maran N. *Project Report for: The Development of a Behavioural Marker System for Newly Qualified Doctors in Managing Acutely Unwell Patients*. Edinburgh; 2013. http://www.docs.hss.ed.ac.uk/iad/Learning_teaching/Academic_teaching/PTAS/Outputs/Mellanby_Jan2012%20award_PTAS_Final_Report.pdf. Accessed March 9, 2019

357. Hull L, Birnbach D, Arora S, Fitzpatrick M, Sevdalis N. Improving surgical ward care: Development and psychometric properties of a global assessment toolkit. *Ann Surg*. 2014;259(5):904-909.

358. Emmert MC, Cai L. A pilot study to test the effectiveness of an innovative interprofessional education assessment strategy. *J Interprof Care*. 2015;29(5):451-456.

359. King HB, Battles J, Baker DP, et al. *TeamSTEPPS™: Team Strategies and Tools to Enhance Performance and Patient Safety*. Agency for Healthcare Research and Quality (US); 2008. https://www.ncbi.nlm.nih.gov/books/NBK43655/?report=reader. Accessed March 9, 2019.

360. Zhang C, Miller C, Volkman K, Meza J, Jones K. Evaluation of the team performance observation tool with targeted behavioral markers in simulation-based interprofessional education. *J Interprof Care*. 2015;29(3):202-208.

361. Andersen PO, Jensen MK, Lippert A, Ostergaard. Development of a formative assessment tool for measurement of performance in multi-professional resuscitation teams. *Resuscitation*. 2010;81(6):703-711.

362. Andrew B, Plachta S, Salud L, Pugh CM. Development and evaluation of a decision-based simulation for assessment of team skills. *Surg (United States)*. 2012;152(2):152-157.

363. Parker-Raley J, Mottet TP, Lawson KA, Duzinski S V, Cerroni A, Mercado M. Investigating pediatric trauma team communication effectiveness phase one: the development of the assessment of pediatric resuscitation communication Investigating pediatric trauma team communication effectiveness phase one: the development of the assessm. *J Commun Healthc*. 2012;5(2):102-115.

364. Parker-Raley J, Yanez K, Cerroni A, Mottet TP, Duzinski S V, Lawson KA. Assessing trauma leader communication in an ED setting. *J Commun Healthc*. 2013;64:197-207.

365. Lambden S, DeMunter C, Dowson A, Cooper M, Gautama S, Sevdalis N. The Imperial Paediatric Emergency Training Toolkit (IPETT) for use in paediatric emergency training: Development and evaluation of feasibility and validity. *Resuscitation*. 2013;84(6):831-836.

366. Sigalet E, Donnon T, Cheng A, et al. Development of a team performance scale to assess undergraduate health professionals. *Acad Med*. 2013;88:989-996.

367. Munroe B, Curtis K, Murphy M, et al. A structured framework improves clinical patient assessment and nontechnical skills of early career emergency nurses: a pre-post study using full immersion simulation. *J Clin Nurs*. 2016;25(15-16):2262-2274.

368. Flynn FM, Sandaker K, Ballangrud R. Aiming for excellence – A simulation-based study on adapting and testing an instrument for developing non-technical skills in Norwegian student nurse anaesthetists. *Nurse Educ Pract*. 2017;22:37-46.

369. Holcomb JB, Dumire RD, Crommett JW, et al. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *J Trauma*. 2002;52(6):1078-85; discussion 1085-6.

370. Rosen MA, Salas E, Wilson KA, et al. Measuring Team Performance in Simulation-Based Training: Adopting Best Practices for Healthcare. *Simul Healthc*. 2008;3(1):33-41.

371. Mitchell L, Flin R, Yule S, et al. Evaluation of the Scrub Practitioners' List of Intraoperative Non-Technical Skills system. *Int J Nurs Stud*. 2012;49(2):201-211.

372. Anderson A, Baxendale B, Scott L, Mossley D. The National Simulation Development Project: Summary Report. http://aspih.org.uk/wp-content/uploads/2017/07/national-scoping-project-summary-report.pdf. Accessed March 9, 2019.

373. Smith-Jentsch KA, Johnston JH, Payne SC. Measuring team-related expertise in complex environments. In: Cannon-Bowers JA, Salas E, eds. *Making Decisions under Stress: Implications*

for Individual and Team Training. Washington: American Psychological Association; 1998:61-87.

374. Wisborg T, Manser T. Assessment of non-technical skills in the operating room - One assessment tool per specialty? *Acta Anaesthesiol Scand*. 2014;58(7):773-774.

375. Cooper S, Porter J, Peach L. Measuring situation awareness in emergency settings: A systematic review of tools and outcomes. *Open Access Emerg Med*. 2013;6:1-7.

376. Shields A, Flin R. Paramedics' non-technical skills: A literature review. *Emerg Med J*. 2013;30(5):350-354.

377. Watanabe Y, Bilgic E, Lebedeva E, et al. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surg Endosc Other Interv Tech*. 2016;30(3):832-844.

378. Onwochei DN, Halpern S, Balki M. Teamwork assessment tools in obstetric emergencies: A systematic review. *Simul Healthc*. 2017;12(3):165-176.

379. Civil Aviation Authority. *Guidance on the Requirements That Pertain to Flightcrew for the Training and Testing of Human Factors Under EASA Part - ORO and EASA Part - FCL*.; 2016. https://publicapps.caa.co.uk/docs/33/CAAStandardsDocument29v7_AUG2016.pdf. Accessed March 9, 2019.

380. Hull L, Arora S, Symons NRA, et al. Training Faculty in Nontechnical Skill Assessment. *Ann Surg*. 2013;258(2):370-375.

381. Robertson ER, Hadi M, Morgan LJ, et al. Oxford NOTECHS II: a modified theatre team non-technical skills scoring system. *PLoS One*. 2014;9(3):e90320.

382. UK Resuscitation Council. Adult advanced life support. https://www.resus.org.uk/resuscitation-guidelines/adult-advanced-life-support/. Published 2015. Accessed March 9, 2019.

383. Jepsen RMHG, Spanager L, Lyk-Jensen HT, Dieckmann P, Østergaard D. Customisation of an instrument to assess anaesthesiologists' non-technical skills. *Int J Med Educ*. 2015;6:17-25.

384. Jepsen RMHG, Ostergaard D, Dieckmann P. Development of instruments for assessment of individuals' and teams' non-technical skills in healthcare: A critical review. *Cogn Technol Work*. 2015;17(1):63-77.

385. Bland J, Altman D. Cronbach's alpha. *Br Med J*. 1997;314:572.

386. Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall; 1991.

387. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

388. Rudolph JW, Simon R, Raemer DB, Eppich WJ. Debriefing as Formative Assessment: Closing Performance Gaps in Medical Education. *Acad Emerg Med*. 2008;15(11):1010-1016.

389. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005;27(1):10-28.

390. Flin R, Patey R. Non-technical skills for anaesthetists: developing and applying ANTS. *Best Pract Res Clin Anaesthesiol*. 2011;25(2):215-227.

391. Gwet K. Benchmarking Inter-Rater Reliability Coefficients. In: *Handbook of Inter-Rater Reliability*. Advanced Analytics LLC; 2014:164-181.

392. Cook DA, Beckman TJ. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *Am J Med*. 2006;119(2):166.e7-166.e16.

393. Guise J-M, Deering SH, Kanki BG, et al. Validation of a tool to measure and promote clinical teamwork. *Simul Healthc.* 2008;3(4):217-223. doi:10.1097/SIH.0b013e31816fdd0a.

394. Lyk-Jensen HT eglgaard, Dieckmann P, Konge L, Jepsen RM alene HG, Spanager L, Østergaard D. Using a Structured Assessment Tool to Evaluate Nontechnical Skills of Nurse Anesthetists. *AANA J*. 2016;84(2):122-127.

395. Baker D, Mulqueen C, Dismukes R. Training raters to assess resource management skills. In: Salas E, Bowers CA, Edens E, eds. *Improving Teamwork in Organizations : Applications of Resource Management Training*. Lawrence Erlbaum; 2001:131-146.

396. Russ S, Hull L, Rout S, Vincent C, Darzi A, Sevdalis N. Observational Teamwork Assessment for Surgery. *Ann Surg*. 2012;255(4):804-809.

397. Yule S, Rowley D, Flin R, et al. Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ J Surg*. 2009;79(3):154-160.

398. Graham J, Hocking G, Giles E. Anaesthesia Non-Technical Skills: Can anaesthetists be trained to reliably use this behavioural marker system in 1 day? *Br J Anaesth*. 2010;104(4):440-445.

399. Patey RE. Identifying and assessing non-technical skills. *Clin Teach*. 2008;5(1):40-41.

400. Huang LC, Conley D, Lipsitz S, et al. The Surgical Safety Checklist and Teamwork Coaching Tools: a study of inter-rater reliability. *BMJ Qual Saf*. 2014;23(8):639-650.

401. Konge L, Vilmann P, Clementsen P, Annema JT, Ringsted C. Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. *Endoscopy*. 2012;44:928-933.

402. Saal F, Downey R, Lahey M. Rating the Ratings: Assessing the Psychometric Quality of Rating Data. *Psychol Bull*. 1980;88(2):413-428.

403. Saner LD, Bolstad CA, Gonzalez C, Cuevas HM. Measuring and predicting shared situation awareness in teams. *J Cogn Eng Decis Mak*. 2009;3(3):280-308.

404. Endsley MR, Rodgers MD. Situation Awareness Information Requirements Analysis for En Route Air Traffic Control. *Proc Hum Factors Ergon Soc Annu Meet*. 1994;38(1):71-75.

405. Riley JM, Kaber DB, Draper J V. Situation awareness and attention allocation measures for quantifying telepresence experiences in teleoperation. *Hum Factors Ergon Manuf*. 2004;14(1):51-67.

406. Prince C, Ellis E, Brannick MT, Salas E. Measurement of Team Situation Awareness in Low Experience Level Aviators. *Int J Aviat Psychol*. 2007;17(1):41-57.

407. Strater LD, Endsley MR, Pleban RJ, Matthews MD, Graham S. *Measures of Platoon Leader Situation Awareness in Virtual Decision-Making Exercises Situation Awareness (SA) Measurement Military Operations on Urbanized Terrain (MOUT) Virtual Environment Infantry Operations Squad Synthetic Environment (SSE).*; 2001. http://www.au.af.mil/au/awc/awcgate/army/rr1770.pdf. Accessed March 9, 2019.

408. Hogan MP, Pace DE, Hapgood J, Boone DC. Use of human patient simulation and the situation awareness global assessment technique in practical trauma skills assessment. *J Trauma*. 2006;61(5):1047-1052.

409. McKenna L, Missen K, Cooper S, Bogossian F, Bucknall T, Cant R. Situation awareness in undergraduate nursing students managing simulated patient deterioration. *Nurse Educ Today*. 2014;34(6):e27-e31.

410. Crozier MS, Ting HY, Boone DC, et al. Use of human patient simulation and validation of the team situation awareness global assessment technique (TSAGAT): A multidisciplinary team assessment tool in trauma education. *J Surg Educ*. 2015;72(1):156-163.

411. Dalkey NC, Helmer O. An experimental application of the Delphi method to the use of experts. *Manage Sci*. 1963;9(3):458-467.

412. UK Resuscitation Council. Peri-arrest arrhythmias. 2015. https://www.resus.org.uk/resuscitation-guidelines/peri-arrest-arrhythmias/. Accessed March 9, 2019.

413. British Thoracic Society. *SIGN 153 • British Guideline on the Management of Asthma A National Clinical Guideline*.; 2016. https://www.sign.ac.uk/assets/sign153.pdf. Accessed March 9, 2019.

414. ESC GUIDELINES 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. https://www.escardio.org/Guidelines/Clinical-Practice-Guidelines/Atrial-Fibrillation-Management. Accessed March 9, 2019

415. National Institute for Health and Care Excellence. Sepsis overview - NICE Pathways. https://pathways.nice.org.uk/pathways/sepsis. Published 2017. Accessed March 9, 2019.

416. American College of Surgeons. Advanced Trauma Life Support. https://www.facs.org/quality-programs/trauma/atls. Accessed March 9, 2019.

417. Cooper S, Kinsman L, Buykx P, McConnell-Henry T, Endacott R, Scholes J. Managing the deteriorating patient in a simulated environment: nursing students' knowledge, skill and situation awareness. *J Clin Nurs*. 2010;19(15-16):2309-2318.

418. Hansel M, Winkelmann AM, Hardt F, et al. Impact of simulator training and crew resource management training on final-year medical students' performance in sepsis resuscitation: A randomized trial. *Minerva Anestesiol*. 2012;78(8):901-909.

419. Crozier MS, Ting HY, Boone DC, O'Regan NB, Bandrauk N. Use of human patient simulation and validation of the team situation awareness global assessment technique (TSAGAT): A multidisciplinary team assessment tool in trauma education. *J Surg Educ*. 2015;72(1):156-163.

420. Chang AL, Dym AA, Venegas-Borsellino C, et al. Comparison between Simulation-based Training and Lecture-based Education in Teaching Situation Awareness A Randomized Controlled Study. *Ann Am Thorac Soc*. 2017;14(4):529-535.

421. Gardner AK, Kosemund M, Martinez J. Examining the Feasibility and Predictive Validity of the SAGAT Tool to Assess Situation Awareness Among Medical Trainees. *Simul Healthc*. 2017;12(1):17-21.

422. Selcon S, Taylor R. Evaluation of the situational awareness rating technique (SART) as a tool for

aircrew systems design. In: *Situational Awareness in Aerospace Operations*. NATO-AGARD; 1990:5/1-5/8.

423. Selcon S, Taylor R, Koritsas E. Workload or situational awareness? In: *Proceedings of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA: Human Factors Society; 1991:62-66.

424. Crabtree M, Marcelo R, McCoy A, Vidulich M. An examination of a subjective situational awareness measure during training on a tactical operations simulator. In: *Proceedings of the Seventh International Symposium on Aviation Psychology*. Columbus, OH: Department of Aviation, The Ohio State University; 1993:891-895.

425. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock PA, Meshkati N, eds. *Human Mental Workload*. North-Holland; 1988:139-183.

426. Vidulich M, Crabtree M, McCoy A. Developing subjective and objective metrics of pilot situation awareness. In: Jensen R, Neumeister D, eds. *Proceedings of the Seventh International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University; 1993:896-900.

427. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika*. 1959;24(2):95-112.

428. Flin R, O'Connor P, Crichton M. Assessing non-technical skills. In: *Safety at the Sharp End*. Boca Raton, FL, US: Ashgate Publishing; 2008:269-306.

429. Endsley MR. Situation Awareness Measurement in Test and Evaluation. In: O'Brien TG, Charlton SG, eds. *Handbook of Human Factors Testing & Evaluation*. ; 1996:159-180.

430. Taylor R, Selcon S. Subjective measurement of situational awareness. In: *Designing for Everyone: Proceedings of the 11th Congress of the International Ergonomics Association*. London: Taylor & Francis; 1991:789-791.

431. Salas E, DiazGranados D, Klein C, et al. Does team training improve team performance? A meta-analysis. *Hum Factors*. 2008;50(6):903-933.

432. Rensink RA. To Have Seen or Not to Have Seen: A Look at Rensink, O'Regan, and Clark (1997). *Perspect Psychol Sci*. 2018;13(2):230-235.

433. Salas E, Prince C, Baker DP, Shrestha L. Situation Awareness in Team Performance: Implications for Measurement and Training. *Hum Factors*. 1995;37(1):123-136.

434. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297-334.

435. Hogg D, Folles K, Strand-Volden F, Torralba B. Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. *Ergonomics*. 1995;38(11):2394-2413.

436. Bolstad CA, Hess TM. Situation awareness and aging. In: *Situation Awareness Analysis and Measurement.* CRC Press; 2000:277-291.

437. Wickens C, Sebok A, Keller J, et al. *Modeling and Evaluating Pilot Performance in NextGen: Review of and Recommendations Regarding Pilot Modeling Efforts, Architectures, and Validation Studies*.; 2013. https://human-factors.arc.nasa.gov/groups/HCSL/publications/HCSL-13-04.pdf. Accessed March 9, 2019.

438. Endacott R, Scholes J, Buykx P, Cooper S, Kinsman L, McConnell-Henry T. Final-year nursing students' ability to assess, detect and act on clinical cues of deterioration in a simulated environment. *J Adv Nurs*. 2010;66(12):2722-2731.

439. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proc Hum Factors Ergon Soc Annu Meet*. 2006;50(9):904-908.

440. Hendy KC, Hamilton KM, Landry LN. Measuring Subjective Workload: When Is One Scale Better Than Many? *Hum Factors*. 1993;35(4):579-601.

441. Liu Y, Wickens CD. Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics. *Ergonomics*. 1994;37(11):1843-1854.

442. Byers, JC. Traditional and Raw Task Load Index (TLX) Correlations: Are paired comparisons necessary? In: Mital A, ed. *Advances in Industrial Ergonomics and Safety*. Taylor & Francis; 1989:481-485.

443. Regehr G, MacRae H, Reznick RK, Szalay D, Redneck R, Scaly D. Comparing the Psychometric Properties of Checklists and Global Rating Scales for Assessing Performance on an OSCE-format Examination. *Acad Med*. 1998;73(9):993-997.

444. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113.

445. Isaak R, Stiegler M, Hobbs G, et al. Comparing Real-time Versus Delayed Video Assessments for

Evaluating ACGME Sub-competency Milestones in Simulated Patient Care Environments. *Cureus*. 2018;10(3):e2267.

446. Driscoll PJ, Paisley AM, Paterson-Brown S. Video assessment of basic surgical trainees' operative skills. *Am J Surg*. 2008;196(2):265-272.

447. Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. *Arthritis Rheum*. 2007;57(5):869-876.

448. Mulcaster JT, Mills J, Hung OR, et al. Laryngoscopic IntubationLearning and Performance. *Anesthesiol J Am Soc Anesthesiol*. 2003;98(1):23-27.

449. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance results of a randomized, double-blinded study. *Ann Surg*. 2002;236(4):458-464.

450. Gwet K. Measures of Association and Item Analysis. In: *Handbook of Inter-Rater Reliability*. Fourth. Advanced Analytics LLC; 2014:343-365.

451. Zheng B, Cassera MA, Martinec D V, et al. Measuring mental workload during the performance of advanced laparoscopic tasks. *Surg Endosc*. 2010;24(1):45-50.

452. Hoonakker P, Carayon P, Gurses AP, et al. IIE Transactions on Healthcare Systems Engineering Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX) *IIE Trans Healthc Syst Eng*. 2011;1:131-143.

453. Gurses AP, Carayon P, Wall M. Impact of Performance Obstacles on Intensive Care Nurses' Workload, Perceived Quality and Safety of Care, and Quality of Working Life. *Health Serv Res*. 2009;44:422-443.

454. Young MS, Brookhuis KA, Wickens CD, Hancock PA. State of science: mental workload in ergonomics. *Ergonomics*. 2015;58(1):1-17.

455. Gaba DM, Lee T. Measuring the Workload of the Anesthesiologist. *Anesth Analg*. 1990;71:354-361.

456. Doyle AC. *The Blanched Soldier*.; 1927. https://sherlock-holm.es/stories/pdf/a4/1-sided/blan.pdf. Accessed March 9, 2019.

457. Vincent C. The Journey to Safety. In: *Patient Safety*. Second. Wiley-Blackwell; 2010:371-404.

458. Health Education England. *National Framework for Simulation Based Education (SBE)*.; 2018. https://www.hee.nhs.uk/sites/default/files/documents/National framework for simulation based education.pdf. Accessed March 9, 2019.

459. Vosper H, Hignett S, Bowie P. Twelve tips for embedding human factors and ergonomics principles in healthcare education. *Med Teach*. 2017;40(4):357-363.

460. Illingworth J. Is the NHS getting safer? 2015. http://www.health.org.uk/sites/health/files/IsTheNHSGettingSafer.pdf. Accessed March 9, 2019.

461. GMC. *Training Environments 2018: Key Findings from the National Training Surveys*.; 2018. https://www.gmc-uk.org/-/media/documents/training-environments-2018_pdf-76667101.pdf?dm_i=2P18,SO0V,1QFRW,2Y9YT,1. Accessed March 9, 2019.

462. The Health Foundation, The King's Fund, The Nuffield Trust. *The Health Care Workforce in England Make or Break?*; 2018. www.hee.nhs.uk/news-blogs-events/news/hee-launches-plan-future-proof-nhs-care-workforce. Accessed March 9, 2019.

463. Lord Darzi. *Better Health and Care for All*.; 2018. https://www.ippr.org/research/publications/the-nhs-long-term-plan. Accessed March 9, 2019.

CHAPTER 11 APPENDICES

APPENDIX 1 NPSA INCIDENT REPORT TEMPLATE: ADAPTED FOR OUHT

Incident Investigation Title:	
Incident Date:	
Incident Number:	
Author(s) and Job Titles	
Investigation Report Date:	

MAIN REPORT:

Incident description and consequences

Incident description:

Incident date:

Actual effect on patient:

Background and context

Add text here

Terms of reference

Guide provided below. Amend this to build your own. Add only a summary to the body of the report.

Purpose To identify the root causes and key learning from an incident and use this information to significantly reduce the likelihood of future harm to patients
Objectives To establish the facts i.e. what happened (*effect*), to whom, when, where, how and why (*root causes*) To establish whether failings occurred in care or treatment To look for improvements rather than to apportion blame To establish how recurrence may be reduced or eliminated To formulate *recommendations and an action plan* To provide a *report and* record of the investigation process & outcome To provide a means of *sharing learning* from the incident To identify routes of *sharing learning* from the incident
Key questions/issues to be addressed

...specific to this incident or incident type
Key Deliverables Investigation Report, Action Plan, Implementation of Actions
Scope (investigation start & end points)
Investigation type, process and methods used Single or Multi-incident investigation Gathering information e.g. *Interviews* Incident Mapping e.g. *Tabular timeline* Identifying Care and service delivery problems e.g. *Change analysis* Identifying contributory factors & root causes e.g. *Fishbone diagrams* Generating solutions e.g. *Barrier analysis*
Investigation team Names, Roles, Qualifications, Departments

Level of investigation

Add text here

Involvement and support of patient and relatives

Add text here

Involvement and support provided for staff involved

Add text here

Information and evidence gathered

Add text here

FINDINGS:

Chronology of events

Chronology (timeline) of events	
Date & Time	Event

Detection of incident

Add text here

Notable practice

Add text here

Care and service delivery problems

Add text here

Contributory factors

Add text here

Root causes

Add text here

Lessons learned

Add text here

CONCLUSIONS:

Recommendations

Add text here

Arrangements for Shared Learning

Add text here

Distribution List

Add text here

Appendices

Add text her

Action Plan

Recommendation	Action	Monitoring of effectiveness	Responsibility	Deadline	Progress Update

APPENDIX 2 NEVER EVENT LIST, NHS ENGLAND 2015-16

Never Event	Description
1. Wrong site surgery	A surgical intervention performed on the wrong patient or wrong site (e.g. wrong knee; wrong organ) includes wrong site nerve block; the incident is detected at any time after the start of the procedure.
2. Wrong implant / prosthesis	Surgical placement of the wrong implant or prosthesis where the implant/prosthesis placed in the patient is other than that specified in the surgical plan either prior to or during the procedure and the incident is detected at any time after the implant/prosthesis is placed in the patient.
3. Retained foreign object post-procedure	Retention of a foreign object in a patient after a surgical/invasive procedure (including interventional radiology, cardiology, interventions related to vaginal birth and interventions performed outside the surgical environment e.g. central line placement in ward areas). Foreign object includes any items that should be subject to a formal counting/checking process before the procedure is completed (e.g. swabs, needles, guide wires etc.).
4. Mis-selection of a strong potassium containing solution	Mis-selection refers to a patient receiving a strong potassium solution intravenously rather than a different, intended medication.
5. Wrong route administration of medication	The patient receives any of the following: • IV chemotherapy intrathecally • Oral/enteral medication or feed/flush administered by any parenteral route • IV administration of a medicine intended for the epidural route.
6. Overdose of insulin due to abbreviations or incorrect device	Overdose refers to: • A patient receiving a tenfold or greater overdose of insulin because a prescriber abbreviates the words 'unit' or 'international units', despite the care setting having an electronic prescribing system in place • A healthcare professional failing to use a specific insulin administration device i.e. does not use an insulin syringe or pen to measure insulin.
7. Overdose of methotrexate for non-cancer treatment	When a patient receives methotrexate, via any route, for non-cancer treatment which results in more than the weekly dose being taken and despite the care setting having an electronic system in place for prescribing and administering/dispensing the medication.

8. Mis-selection of high strength midazolam during conscious sedation	Mis-selection means that a patient receives an overdose due to the selection of a high strength midazolam preparation (5 or 2 mg/ml) rather than the 1 mg/ml preparation, in a clinical area performing conscious sedation.
9. Failure to install functional collapsible shower or curtain rails	• Failure of collapsible curtain or shower rails to collapse when an inpatient suicide is attempted/successful • Failure to install collapsible rails and an inpatient suicide is attempted/successful using these non-collapsible rails.
10. Falls from poorly restricted windows	• Applies to windows 'within reach' of patients. This means windows within reach of someone standing at floor level and that can be exited/fallen from without the need to move furniture or use tools to assist climbing out of the window. • Includes windows in facilities/areas where healthcare is provided and where patients can and do access • Includes where patients deliberately or accidentally fall from a window where restrictor has been fitted but previously damaged or disabled • Includes where patients are deliberately able to overcome a window restrictor by hand
11. Chest or neck entrapment in bedrails	Entrapment of a patient's chest or neck within or between bedrails, bedframe and mattress where bedrail dimensions do not comply with Medicines and Healthcare products Regulatory Agency guidance.
12. Transfusion or transplantation of ABO-incompatible blood components or organs	Unintentional transfusion of ABO-incompatible blood transfusion or unintentional ABO mismatched transplantation of solid organs
13. Misplaced naso- or oro-gastric tubes	Misplacement and use of a naso- or oro-gastric tube in the pleura or respiratory tract where the misplacement of the tube is not detected prior to commencement of feeding, flush or medication administration.
14. Scalding of patients	Patient being scalded by water used for washing/bathing.

APPENDIX 3 CONTRIBUTORY FACTORS FROM THE LONDON PROTOCOL

Factor Types	Specific factors contributing to incident
Patient Factors	Condition (complexity & seriousness) Language and communication Personality and social factors
Task and Technology	Task design and clarity of structure Availability and use of protocols Availability and accuracy of test results Decision-making aids
Individual (staff)	Knowledge and skills Competence Physical and mental health
Team	Verbal communication Written communication Supervision and seeking help Team structure (congruence, consistency, leadership, etc)
Work / Environmental:	Staffing levels and skills mix Workload and shift patterns Design, availability and maintenance of equipment Administrative and managerial support Environment Physical
Organisational / Management	Financial resources & constraints Organisational structure Policy, standards and goals Safety culture and priorities
Institutional Context:	Economic and regulatory context National health service executive Links with external organisations

APPENDIX 4 NCEPOD ACUITY CODES

Acuity code and NCEPOD category	Definition
1 - Elective	Intervention planned or booked in advance of routine admission to hospital. Timing to suit patient, hospital and staff.
2 - Expedited	Patient requiring early treatment where the condition is not an immediate threat to life, limb or organ survival. Normally within days of decision to operate.
3 - Urgent	Intervention for acute onset or clinical deterioration of potentially life threatening conditions, for those conditions that may threaten the survival of limb or organ, for fixation of many fractures and for relief of pain or other distressing symptoms. Normally within hours of the decision to operate.
4 - Immediate	Immediate life, limb or organ-saving intervention – resuscitation simultaneous with intervention. Normally within minutes of decision to operate.

APPENDIX 5 EXAMPLES OF MEDICAL CASES WITH ACUITY CODES ALIGNED TO NCEPOD CLASSIFICATION

Acuity code (explanation)	Example of medical SIRI (case index no.)
1 – Elective (patient in hospital for a procedure which has been arranged at his convenience)	Patient admitted for elective procedure (thermal ablation of liver lesion) with an overnight stay (normal practice) suffers a fall (105)
2 – Expedited (medical care required on a daily basis until cause of falls can be ascertained)	Elderly patient admitted with history of dementia and repeated falls at home develops a pressure ulcer on his heel during his admission (160)
3 – Urgent (medical care of poorly controlled diabetes with associated infection should be on hourly basis)	Patient with poorly controlled diabetes seen in emergency department at Horton Hospital with infected foot, incorrect decision to treat him daily as an outpatient (109)
4 – Immediate (medical care of sepsis requires immediate provision of appropriate intravenous antibiotics)	Patient recovering from sub-arachnoid haemorrhage develops severe urosepsis at the Oxford Centre for Enablement and treatment is delayed (72)

APPENDIX 6 DEFINITION OF CLINICAL CATEGORY WHERE MORE THAN ONE CATEGORY WAS ASSIGNED (UNDERLINED TEXT SHOWS FINAL, AGREED ALLOCATION)

Incident summary (case index no.)	**Incident Category**					
	Sepsis	Delayed or incorrect diagnosis	Diabetes management	Medication error	Foreign body	Device
Pericardial and pleural effusions in a neonate, caused by long line damage to heart. (7)		***Delayed diagnosis of pericardial and pleural effusions.***				Inadequate SOPs for use of long line in neonates. Inadequate light source for viewing X-rays in neonatal unit.
Patient has steroid implant (for delayed release of drug) inserted into an incorrect position in the eye. (56)				Steroid in the wrong compartment of the eye.		***Bespoke delivery device not used by this surgeon before.***
Patient with infected foot suffers cardiac arrest shortly after arrival on ward. (61)	***Septic foot diagnosed promptly in outpatients but inadequately treated.***			Delay in prescribing and giving intravenous antibiotics.		
Patient who had suffered an intracerebral bleed developed urosepsis for which treatment was delayed. (72)	***Urosepsis diagnosed in rehab centre where patient is recovering well but is not able to communicate clearly***			Transferred to JR for treatment and antibiotics are delayed and the incorrect dose is given		

Patient is incorrectly diagnosed and treated for PE and then suffers a bleed from her bowel. (90)		***Incorrect diagnosis of PE from a CT scan.***		Anticoagulant treatment prescribed on basis of incorrect diagnosis.		
Contrast for a scan of the head is incorrectly injected into the port of a catheter lying in the brain instead of the vein. (97)				***Incorrect site of injection of contrast medium (gadolinium).***		Two similar devices in close proximity both with red caps, one in a large vein and one in the brain.
Patient discharged from hospital with two types of insulin instead of one, has hypoglycaemic event. (98)			Inappropriate management of diabetes – excess insulin inadvertently given by district nurse.	***Incorrect prescription of take home medication not detected. Patient given too much insulin which led to hypoglycaemic event.***		
Avoidable DKA in a patient who has a subdural haematoma. (99)		Delayed diagnosis of DKA: symptoms masked by intracerebral bleed.	***Inadequate monitoring of diabetes led to failure to detect high blood sugars.***			
Patient with diabetes and infected foot requires emergency surgical intervention to amputate. (109)	Patient initially has isolated infection of foot which progresses to sepsis due incorrect risk stratification.	***Delayed diagnosis of severity of infection led to initial management as an outpatient.***	Inadequate management of DM, lack of recognition that diabetes is an additional risk factor for foot infection.			

Patient with sepsis suffers MI which is not recognized. (133)	Patient admitted with sepsis which is treated but symptoms and signs of MI not detected.	***Delayed diagnosis of MI in a patient who was triggering on the SEND system.***				
Patient admitted for treatment to control Diabetes is discharged with a cannula in his arm which became infected. (142)	Patient admitted with sepsis secondary to cannula left in situ after discharge from hospital.		Patient admitted for treatment of diabetes.		***Cannula left in situ and not detected on discharge from ward.***	
Patient admitted with diabetes and hyperglycaemia. (149)			***Inadequate monitoring and treatment of diabetes led to worsening blood sugar levels.***	Inadequate insulin given.		
Patient seen in ambulatory unit with diabetes, sent home with inadequate treatment and returned with DKA. (153)			***Inadequate management of diabetes.***	Failure to use insulin in a patient with previously diagnosed type 2 diabetes.		

APPENDIX 7 SEARCH CRITERIS FOR SYSTEMATIC REVIEW OF TOOLS FOR THE ASSESSSMENT OF NTS IN HEALTHCARE

SEARCH CRITERIA

PROTOCOL AND REGISTRATION

This systematic review was registered with Prospero ref. no: CRD42017055445 and can be found at https://www.crd.york.ac.uk/prospero/ .

ELIGIBILITY CRITERIA AND INFORMATION SOURCES

Peer-reviewed studies were identified by search of the electronic bibliographic databases Medline; Embase; CINAHL; PsycINFO; Scopus and ERIC. A search of the grey literature was made via Google Scholar, ProQuest and OpenGrey. A manual search of the reference list of identified relevant articles was also conducted.

Papers were limited to publication after 1990 (the year that the first techniques for use in aviation was described) and English language.

All reviewed articles had their quality assessed using criteria defined by the Hawker assessment tool for mixed methods research studies as per guidance from Prospero https://www.crd.york.ac.uk/prospero/#aboutpage . The assessment form adapted from Hawker's guidance and used by the authors in deciding on inclusion or rejection of a study is included below.

The electronic search strategy used was:

- Non-technical skills assessment tools:

  (Non-technical skill* OR behavio* rating system OR behavio* marking system OR human factors OR ergonomics) AND (assessment OR tool*) AND (healthcare OR operating theatre OR operating room OR intensive care unit OR critical care unit OR emergency department OR trauma OR ward OR hospital OR surgery) OR (simulat* centre OR simulat* scenario OR simulat* environment) AND (situatio*awareness OR communicat* OR teamwork* OR task management OR task allocation OR decision making)

- Named tools: Once named tools were found further searches were undertaken using the tool's name to uncover any additional scoring systems and to look for further evidence of validity and reliability.

INCLUSION CRITERIA

Papers were eligible for inclusion where:

- They were published in the English language, or translation was available
- The population studied comprised healthy adults working in healthcare settings
- The publication date was between January 1990 and March 2017
- They described a tool designed to assess non-technical skills and included more than one of the following domains: communication, teamwork, situation awareness, decision making and task allocation/management.
- They described a tool designed for use by direct observation or review of audio-visual files in a simulated or real clinical setting.

EXCLUSION CRITERIA

Papers were excluded where:

- Reports had not undergone peer-review
- Ethical approval of the study or informed consent from participants was not described
- No data on the tool's validity or reliability were available
- The tool was designed for self-assessment only.
- The tool did not analyse performance under more than one of the key non-technical domains of: communication, situation awareness (sometimes described as vigilance), decision making or task allocation/management
- They described a tool used for the study of technical skills only

DATA ITEMS AND SUMMARY MEASURES

Data were abstracted using the following checklists:

- **Trial objective:** short descriptive statement outlining the intent of the study
- **Study type:** observational, single- or repeated measures
- **Study setting:** undergraduates, or descriptor of study population and (where relevant) description of participant group
- **Study location:** the geographic region and environment (clinical or training) in which the study was carried out
- **Study participant demographics:** age and healthcare background of participants in each study
- **Sample size enrolled:** the number of participants initially recruited into the study, whether they were included in the reported analysis

- **Evidence of the validity of the tool:** Construct validity and sub-classifications (face, content, discriminant, concurrent, convergent and predictive validity) and response process for the tool described in the study: e.g. description of type of scoring system, a justification for its use or an explanation of how assessors were trained in its use and comparison with other systems
- **Evidence of the reliability of the tool:** intra- and inter- operator reliability and (where reported) consistency between training centres, internal consistency
- **Measure of ease of use of the tool:** either qualitative or quantitative
- **Measure of the training required to use the tool or the cost of using the tool**

APPENDIX 8 ASSESSMENT QUESTIONNAIRE FOR PAPERS IN SYSTEMATIC REVIEW

Assessment questionnaire

Author(s): Date of Publication:

Abbreviated Title: Reviewer:

RELEVANCE TO RESEARCH QUESTIONS

[] Has the tool been designed for use in assessing healthcare professionals?

[Sim] [Real] [Both] Is the tool for use in simulated or real clinical settings?

[] Does the tool describe non-technical skills domains?

How has the tool been designed?

[] Literature review [] Delphi type process with subject matter experts [] Formal task analysis

[] Has the tool been assessed for validity?

[] Has the tool been assessed for reliability?

NTS DOMAINS SCORED

[] Communication

[] Leadership/teamwork (together or separately)

[] Situation Awareness (or equivalent e.g. vigilance)

[] Decision making

[] Task allocation/management

[] Other (if deemed relevant)

VALIDITY

Has evidence of construct validity for the tool been provided in any of the following forms?

[] Face validity

[] Content validity

[] Discriminant validity

[] Convergent validity

[] Concurrent validity

[] Predictive validity

RELIABILITY

Has the tool been analysed for reliability in any of the following forms?

[] Internal consistency

[] Inter-rater reliability

[] Intra-rater reliability

[] Test-retest reliability

USABILITY

Has the usability of the tool been assessed?

[] usability described qualitatively

[] usability described quantitatively

[] training requirements described

STUDY TYPE

[] Empirical study (peer reviewed)

[] Empirical study (non-peer reviewed)

[] Assessment tool designed by other agency (e.g. healthcare organisation or charitable body)

[] Other

APPENDIX 9 ACRONYMS FOR TOOLS FOR THE ASSESSMENT OF NTS IN HEALTHCARE

NTS Assessment tool Acronym	Full name
AeroNOTS	Aeromedical Non-Technical Skills
Anaesthetic trainee NTS	Unnamed tool for assessment of NTS in anaesthetic trainees
ANTS	Anaesthetists' Non-Technical Skills
ANTS-AP	Anaesthetic Non-Technical Skills for Anaesthetic Practioners
ANTSdk	Anaesthesiologists' Non-Technical Skills in Denmark
AOTP / GAOTP	Assessment of Obstetric Team Performance / Global AOTP
APRC	Assessment of Paediatric Resuscitation Communication
BAR ACRM	Behaviourally Anchored Rating scale in Anesthesia Crisis Resource Management
BMS-NNTS	Behavioural Marker System – Neurosurgical Non-Technical Skills
BRS-ECMO	Behavioral Rating Score for Extra-Corporeal Membrane Oxygenation
CARDIOTEAM	CARDIOTEAM course assessment tool
CATS	Communication and Teamwork Skills
CBOATs	Collaborative Behaviors Objective Assessment Tools
CEA	Checklist of Expected Actions – obstetric crises
CTS	Clinical Teamwork Scale
Emergency Dr NTS	Unnamed tool for assessment of NTS in emergency medicine doctors
Emergency team NTS	Unnamed tool for assessment of NTS in Emergency teams
Endo-OTAS	Endovascular Observational Teamwork Assessment for Surgery
ENNTS	Emergency Nurses' Non-Technical Skills
EPOC	Explicit Professional Oral Communication
F -TEAM	French - Team Emergency Assessment Measure
FoNTS	Foundation Non-Technical Skills
GRS-A	Global Rating Scale for Anaesthetists
Healthcare students NTS	Unnamed tool for assessment of NTS in healthcare students
HPAT	Human Performance Assessment Tool
ICARS	Interpersonal and Cognitive Assessment for Robotic Surgery
IMCBRS	Immediate Medical Care Behaviour Rating System
IMPAcT	Imperial Military Personnel Assessment (c) Tool
IPETT	Imperial Paediatric Emergency Training Toolkit
iTOFT	individual Teamwork Observation and Feedback Tool
KidSIM	Kid Simulation – Team Performance Scale
L&TBM	Leadership and Team Behavior Measurement tool
MHPTS	Mayo High Performance Teamwork Scale
MINTS-DR	Multi-professional Inventory for Non-Technical Skills in the Delivery Room
NANTS-no	Nurse Anaesthetists' Non-Technical Skills in Norway
NANTSdk	Nurse Anaesthetists' Non-Technical Skills in Denmark
NOTSS	NOn-Technical Skills for Surgeons
NOTSSdk	NOn-Technical Skills for Surgeons in Denmark
NTS sepsis	Unnamed tool for assessment of NTS in sepsis
ORCA	Operating Room Communication Assessment

ORTAS	Operating Room Teamwork Assessment Scale
OSANTS	Objective Structured Assessment of Non-Technical Skills
OSCAR	Observational Skill-based Clinical Assessment tool for Resuscitation
OTAS	Observational Teamwork Assessment for Surgery
OTAS - WR	Observational Teamwork Assessment for Surgery for ward rounds
OTAS-D	Observational Teamwork Assessment for Surgery - Deutsch
OTAS-S	Observational Teamwork Assessment for Surgery - Spanish
OTRS	Observer Teamwork Rating Scale
Ottawa GRS	Ottawa crisis resource management Global Rating Scale
Ottawa GRS-I	Ottawa crisis resource management Global Rating Scale - Italy
Oxford NOTECHS	Oxford NOn-TECHnical Skills
Paramedic NTS	Unnamed tool for assessment of NTS in Paramedics
PCST	Paediatric Cardiac Surgery Teamwork score
PETRA	Perinatal Emergency Team Response Assessment
PR-NTS	Pediatric Resident – Non-Technical Skills
PS-OSCE-NTS	Procedural Skills – Objective Structured Clinical Examination – Non-Technical Skills
Revised NOTECHS	Revised NOn-TECHnical Skills scale
SAFE-TeamS	Standardised Assessment for Evaluation of Team Skills
SPLINTS	Scrub Practitioners List of Intraoperative Non-Technical Skills
SSC and TCT	Surgical Safety Checklist and Teamwork Coaching Tools
STAT	Simulation Team Assessment Tool
Surgical NTS	Unnamed tool for assessment of NTS in Surgery
SWAT	Surgical Ward round Assessment Tool
TLIS	Teamwork Leadership Interpersonal Skills
T-MEX	Teamwork –Mini-clinical evaluation Exercise
T-SAW-C	Teamwork Skills Assessment for Ward Care
TBR	Teamwork Behavioural Rater
TDFR	Team Dimensions Rating Form
TEAM	Team Emergency Assessment Measure
T&PCM	Teamwork and Patient Care Measure
TFAT	Team Functioning Assessment Tool
TPOT	Team Performance Observation Tool
Trauma NOTECHS	Trauma NOn-TECHnical Skills scale
Trauma team NTS	Unnamed tool for assessment of NTS in Trauma Teams
TTCA-24	Trauma Team Communication Assessment - 24
UTBMNR	University of Texas Behavioral Markers for Neonatal Resuscitation
WHOBARS	WHO checklist Behaviorally Anchored Rating Scale

N.B. Tools are in alphabetical order

APPENDIX 10 COMPLETE TABLE OF ATTRIBUTES OF ALL 76 TOOLS FOR THE ASSESSMENT OF NTS IN HEALTHCARE

Tool (Year)	Environment		Specialty / clinical arena	MDT involved plus Psych/HF	Validity		Reliability		Usability		Total score (max 27)
	Sim	Real			Response process	>2 items	internal consistency	Interrater reliability	Usability evaluated	Training specified	
TEAM (2010)	●	●	Emergency Department MDT	●	●	●	●	●	●	●	25
OTAS (2004)	●	●	Theatre MDT	●		●	●	●	●	●	24
Oxford NOTECHS (2008)	●	●	Theatre MDT	●	●	●	●	●		●	24
ANTS (2003)	●	●	Anaesthetists	●	●		●	●	●	●	23
Ottawa GRS (2006)	●		Medical trainees (any specialty)	●	●	●	●	●	●	●	23
NOTSS (2006)	●	●	Surgeons	●	●		●	●	●	●	22
ANTS-AP (2015)	●	●	Anaesthetic practitioners	●			●	●	●	●	22
PETRA (2017)	●	●	Obstetric MDT		●	●	●	●	●	●	22
UTBMNR (2004)	●	●	Neonatal MDT	●	●		●	●		●	21
AOTP/GAOTP (2009)	●		Obstetric MDT	●		●	●	●	●	●	21
PCST (2010)		●	Paediatric cardiac surgery MDT	●	●		●	●	●	●	21
SPLINTS (2011)	●	●	Scrub practitioners		●		●	●	●	●	21
TFAT (2011)	●	●	Ward MDT	●		●		●	●	●	21
OSCAR (2011)	●	●	Resuscitation MDT	●	●		●	●		●	21
WHOBARS (2016)	●	●	Theatre MDT	●			●	●	●	●	21
CTS[311] (2008)	●	●	Healthcare MDT	●	●			●	●	●	20
SAFE-TeamS (2013)	●		Students (medical, nursing)	●	●		●	●	●	●	20
ANTSdk (2015)	●		Anaesthetists	●	●		●	●	●	●	20
TDRF (2002)		●	Emergency Department MDT	●		●	●	●		●	19
Revised NOTECHS (2005)	●		Theatre MDT	●		●	●	●		●	19
MHPTS* (2007)	●		Healthcare MDT	●			●	●	●	●	19
T&PCM (2008)	●		Emergency Department MDT	●		●		●		●	19
Anaesthetic trainee NTS (2010)	●	●	Anaesthetic trainees			●	●	●		●	19
Trauma NOTECHS (2012)	●	●	Trauma MDT			●	●	●		●	19
NANTSdk (2014)	●	●	Nurse anaesthetists	●			●	●	●	●	19

T-MEX (2014)		●	Students (medical) or trainee doctors				●	●	●	●	19
SWAT (2015)	●	●	Surgeons on ward rounds	●		●		●	●		19
OSANTS (2015)	●	●	Surgical trainees		●		●	●			19
Endo-OTAS (2016)	●	●	Endovascular MDT	●					●	●	19
AeroNOTS (2016)	●		Doctors in aeromedical transport		●	●		●		●	19
iTOFT (2016)	●	●	Students (nursing, AHP)				●		●	●	19
Emergency Dr NTS (2011)		●	Emergency Medicine doctors	●				●		●	18
TBR (2011)	●		Intensive care MDT	●		●	●			●	18
NOTSSdk (2012)	●	●	Surgeons	●			●	●		●	18
FoNTS (2013)	●	●	Foundation doctors	●			●			●	18
T-SAW-C (2014)	●	●	Surgical trainees on ward rounds	●			●	●		●	18
BMS-NNTS (2014)		●	Neurosurgeons	●	●			●		●	18
Healthcare Student NTS (2015)	●		Students (AHP)	●			●	●	●		18
ICARS (2017)	●		Surgeons in robotic surgery			●	●	●	●		18
TPOT (2008)	●		Healthcare MDT	●			●	●		●	17
CARDIOTEAM (2010)	●		Resuscitation MDT		●			●	●	●	17
Surgical NTS (2012)	●		Surgical trainees			●	●	●			17
APRC (2012)		●	Trauma MDT					●	●	●	17
IPETT (2013)	●		Paediatric and anaesthetic trainees	●		●	●			●	17
KidSIM (2013)	●		Students (medical, nursing, AHP)				●	●		●	17
OTAS-S (2014)		●	Theatre MDT	●				●		●	17
ENNTS (2016)	●		Emergency Department nurses		●	●	●			●	17
MINTS-DR (2017)	●		Obstetric MDT	●					●	●	17
NANTS-no (2017)	●	●	Nurse anaesthetists				●	●		●	17
HPAT (2002)	●		Trauma MDT		●	●			●		16
GRS-A (2003)	●		Anaesthetists			●	●	●		●	16
CATS (2007)	●	●	Healthcare MDT	●	●					●	16
PR-NTS (2012)	●		Paediatric residents			●	●	●			16
OTAS - WR (2012)		●	Medical ward round MDT	●				●		●	16
OTAS-D (2013)	●	●	Theatre MDT	●				●		●	16

SSC and TCT (2014)		●	Theatre MDT					●	●	●	16
PS-OSCE-NTS (2014)	●		General medicine trainees		●	●	●			●	16
Ottawa GRS-I[283] **(2016)**	●		Medical trainees (any specialty)			●	●	●		●	16
IMCBRS (2017)	●		Pre-hospital MDT	●				●	●		16
BAR ACRM (1998)	●		Anaesthetists	●				●		●	15
NTS sepsis (2007)	●		Intensive Care MDT			●		●		●	15
L&TBM (2009)	●		Students (medical)				●	●		●	15
Trauma NTS (2009)	●		Trauma surgeons		●			●			15
EPOC (2013)		●	Emergency Department and Intensive Care MDT	●				●		●	15
CBOATs (2014)	●		Students (medical, nursing)				●	●		●	15
ORTAS (2014)	●		Students (medical, nursing, AHP)			●		●		●	15
F-TEAM (2016)	●		Emergency Department MDT	●			●	●		●	15
TTCA-24 (2017)		●	Trauma MDT			●		●		●	15
BRS-ECMO (2006)	●		ECMO nurses				●	●			14
OTRS (2009)	●		Students (medical)	●				●		●	14
STAT (2012)	●		Paediatric residents					●		●	14
TLIS (2014)	●		Intensive Care doctors / advanced practitioners			●		●			14
Paramedic NTS (2008)	●		Paramedics				●	●			13
CEA (2008)	●		Obstetric trainees					●		●	12
ORCA (2012)	●		General surgeons and Gynaecologists								12
Emergency team NTS (2016)	●		Emergency Department MDT					●			11

AHP – Allied Health Professions
MDT – multidisciplinary team (this was marked on the table if more than one professional group plus additional expertise from psychologists (Psych) or human factors (HF) were involved in the design of the tool)

APPENDIX 11 ADDITIONAL INFORMATION ON SCORES FOR NTS ASSESSMENT TOOL DEVELOPMENT, PSYCHOMETRIC TESTING, NTS CATEGORIES AND COUNTRY OF ORIGIN

Tool (Year)	Country of origin	Specialty / clinical arena	Communication	Leadership/ teamwork	Situation awareness	Decision making	Task management management	Development score (max 15)	Psychometric score (max 12)	Total (max 27)
TEAM (2010)	Australia	Emergency Department MDT	●	●	●		●	15	10	25
OTAS (2004)	UK	Theatre MDT	●	●	●		●	14	10	24
Oxford NOTECHS (2008)	UK	Theatre MDT	●	●	●	●	●	14	10	24
ANTS (2003)	UK	Anaesthetists	●	●	●	●	●	14	9	23
Ottawa GRS (2006)	Canada	Medical trainees (any specialty)	●	●	●	●	●	13	10	23
NOTSS (2006)	UK	Surgeons	●	●	●	●	●	14	8	22
ANTS-AP (2015)	UK	Anaesthetic practitioners	●	●	●		●	13	9	22
PETRA (2017)	Canada	Obstetric MDT	●	●	●		●	13	9	22
UTBMNR (2004)	USA	Neonatal MDT	●	●	●		●	15	6	21
AOTP/GAOTP (2009)	Canada	Obstetric MDT	●	●	●		●	11	10	21
PCST (2010)	Netherlands	Paediatric cardiac surgery MDT	●	●	●	●	●	14	7	21
SPLINTS (2011)	UK	Scrub practitioners	●	●	●		●	13	8	21
TFAT (2011)	Australia	Ward MDT	●	●	●		●	13	8	21
OSCAR (2011)	UK	Resuscitation MDT	●	●	●	●	●	15	6	21
WHOBARS (2016)	New Zealand	Theatre MDT	●	●	●			13	8	21
CTS (2008)	USA	Healthcare MDT	●	●	●	●		15	5	20
SAFE-TeamS (2013)	USA	Students (medical, nursing)	●	●	●			12	8	20
ANTSdk (2015)	Denmark	Anaesthetists	●	●	●	●	●	12	8	20
TDRF (2002)	USA	Emergency Department MDT	●	●	●	●	●	12	7	19
Revised NOTECHS (2005)	UK	Theatre MDT	●	●	●	●	●	12	7	19
MHPTS* (2007)	USA	Healthcare MDT	●	●	●		●	12	7	19
T&PCM (2008)	USA	Emergency Department MDT	●	●	●	●	●	12	7	19
Anaesthetic trainee NTS (2010)	UK	Anaesthetic trainees	●	●	●	●	●	12	7	19
Trauma NOTECHS (2012)	USA	Trauma MDT	●	●	●	●	●	12	7	19
NANTSdk (2014)	Denmark	Nurse anaesthetists	●	●	●	●	●	12	7	19
T-MEX (2014)	Australia	Students (medical) or trainee doctors	●	●				10	9	19
SWAT (2015)	UK	Surgeons on ward rounds	●	●	●	●		12	7	19
OSANTS (2015)	Canada	Surgical trainees	●	●	●	●		13	6	19
Endo-OTAS (2016)	UK	Endovascular MDT	●	●	●		●	13	6	19
AeroNOTS (2016)	New Zealand	Doctors in aeromedical transport	●	●	●	●	●	11	8	19
iTOFT (2016)	Australia	Students (nursing, AHP)	●	●	●	●	●	13	6	19
Emergency Dr NTS (2011)	UK	Emergency Medicine doctors	●	●	●	●	●	11	7	18

TBR (2011)	New Zealand	Intensive care MDT	●	●	●		●	11	7	18
NOTSSdk (2012)	Denmark	Surgeons	●	●	●	●		13	5	18
FoNTS§ (2013)	UK	Foundation doctors	●	●	●	●	●	12	6	18
T-SAW-C (2014)	UK	Surgical trainees on ward rounds	●	●	●	●	●	12	6	18
BMS-NNTS (2014)	France	Neurosurgeons	●	●	●	●	●	13	5	18
Healthcare Student NTS (2015)	USA	Students (AHP)	●	●	●			10	8	18
ICARS (2017)	UK	Surgeons in robotic surgery	●	●	●	●	●	8	10	18
TPOT† (2008)	USA	Healthcare MDT	●	●	●		●	10	7	17
CARDIOTEAM (2010)	Denmark	Resuscitation MDT	●	●	●		●	12	5	17
Surgical NTS (2012)	USA	Surgical trainees	●	●	●	●	●	10	7	17
APRC (2012)	USA	Trauma MDT	●	●			●	11	6	17
IPETT (2013)	UK	Paediatric and anaesthetic trainees	●	●		●	●	11	6	17
KidSIM (2013)	Canada	Students (medical, nursing, AHP)	●	●	●		●	11	6	17
OTAS-S (2014)	Colombia	Theatre MDT	●	●	●		●	13	4	17
ENNTS (2016)	Australia	Emergency Department nurses	●		●	●	●	11	6	17
MINTS-DR (2017)	Italy	Obstetric MDT	●	●	●	●	●	12	5	17
NANTS-no (2017)	Norway	Nurse anaesthetists	●	●	●	●	●	10	7	17
HPAT (2002)	USA	Trauma MDT	●	●				10	6	16
GRS-A (2003)	UK	Anaesthetists	●	●	●		●	8	8	16
CATS (2007)	USA	Healthcare MDT	●	●	●		●	13	3	16
PR-NTS (2012)	Canada	Paediatric residents	●	●	●		●	10	6	16
OTAS - WR (2012)	USA	Medical ward round MDT	●	●	●		●	11	5	16
OTAS-D (2013)	Germany	Theatre MDT	●	●	●		●	13	3	16
SSC and TCT (2014)	USA	Theatre MDT	●	●			●	11	5	16
PS-OSCE-NTS (2014)	Canada	General medicine trainees	●	●			●	11	5	16
Ottawa GRS-I (2016)	Italy	Medical trainees (any specialty)	●	●	●	●	●	9	7	16
IMCBRS (2017)	UK	Pre-hospital MDT	●	●	●	●		11	5	16
BAR ACRM (1998)	USA	Anaesthetists	●	●	●		●	10	5	15
NTS sepsis (2007)	USA	Intensive Care MDT	●	●	●		●	10	5	15
L&TBM (2009)	USA	Students (medical)	●	●	●		●	9	6	15
Trauma NTS (2009)	USA	Trauma surgeons	●	●	●		●	11	4	15
EPOC (2013)	Netherlands	Emergency Department and Intensive Care MDT	●	●	●		●	12	3	15
CBOATs (2014)	USA	Students (medical, nursing)	●			●		9	6	15
ORTAS (2014)	USA	Students (medical, nursing, AHP)	●	●	●		●	10	5	15
F-TEAM (2016)	France	Emergency Department MDT	●	●	●		●	9	6	15
TTCA-24 (2017)	USA	Trauma MDT	●	●			●	10	5	15
BRS-ECMO (2006)	USA	ECMO nurses	●	●	●		●	9	5	14
OTRS (2009)	USA	Students (medical)	●	●	●	●		10	4	14
STAT (2012)	USA	Paediatric residents	●	●	●	●	●	9	5	14
TLIS (2014)	USA	Intensive Care doctors/ advanced practitioners	●	●	●	●		10	4	14

Paramedic NTS (2008)	Switzerland	Paramedics	●	●		●	●	9	4	13
CEA (2008)	USA	Obstetric trainees	●	●	●		●	9	3	12
ORCA (2012)	USA	General surgeons and Gynaecologists	●	●	●			9	3	12
Emergency team NTS (2016)	Israel	Emergency Department MDT	●	●				7	4	11

APPENDIX 12 FACULTY GUIDELINES FOR CARDIAC ARREST SCENARIO USED FOR ASSESSMENT OF NTS TOOLS IN CHAPTER 6

Scenario summary and main learning objectives:

1) ABCDE assessment of critically ill patient

2) Recognition of PEA and management of cardiac arrest with a reversible cause

PMH: Asthma

DH: Salbutamol, beclomethasone

Allergies: Intolerant of NSAIDs

Scenario Time Point	Sim Man Settings	Expected candidate actions
START	A Patent B Widespread wheeze, SpO2 60% on air (rise to 86% on O2), RR 40 Pneumothorax C Tachycardic on presentation	Competent ABCDE approach to critical illness Give O2 Consider appropriate drug management of acute severe asthma.
MIDDLE	Starts when select next palette item PEA arrest (sinus tachycardia)	Management of PEA arrest according to ALS algorithm Adrenaline during 1st cycle Consideration and recognition of reversible causes Needle decompression (if not already done) Good team leadership skills
END	Moves to when select next palette. ROSC after 2nd shock.	Appropriate shock delivery Confirm ROSC and post-resus care

Scenario Number: StART Assessment(ALS)

Name: Andrew Jones

Age: 25
Hospital Number: 1234567
Location: Emergency Department; Resus Bay

Student brief: Mr Jones is a 25 year old man with asthma who has become acutely short of breath this morning. He has taken his normal salbutamol with no effect, and has been admitted by ambulance to the emergency department.

Clinical Case: Mr Jones presents in extremis with acute severe asthma, with life-threatening signs. At some point during the ABCDE assessment he collapses in PEA. If students consider reversible causes, they will find a left sided pneumothorax.

History: Mr Jones has a lifelong history of asthma and was ventilated for a period of 8 days 6 months ago. He normally takes salbutamol (which he uses daily) and beclomethasone.

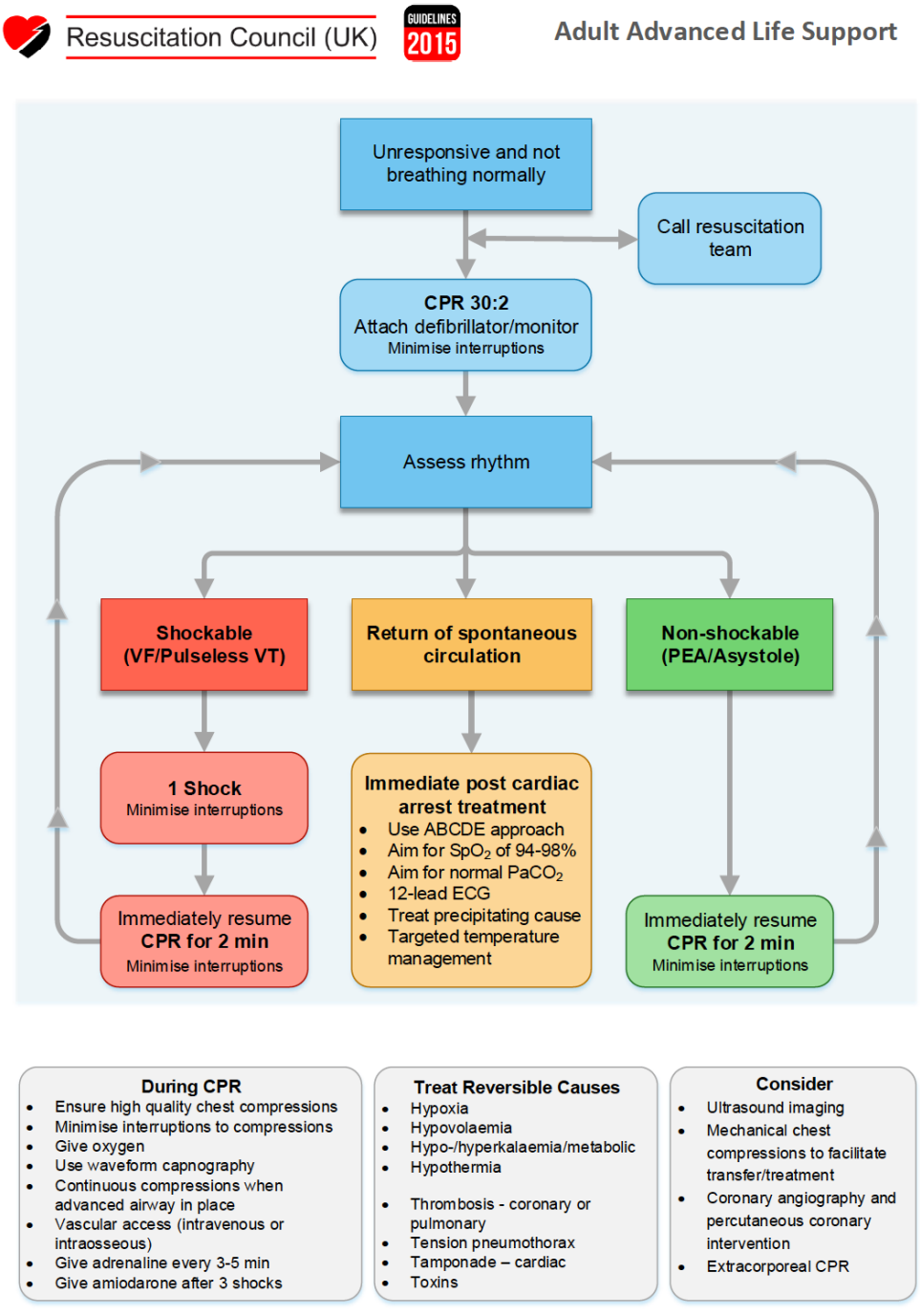
Investigations: ABG reveals type II respiratory failure. Chest X-ray is not available.

.

Instructions for Faculty:

Set Up:
Manikin with nebuliser running
Manikin in sitting position

APPENDIX 13 RESUSCITATION COUNCIL ADVANCED LIFE SUPPORT GUIDELINES



APPENDIX 14 ANTS SCORE SHEET

Category	Element	*Rating	Observation on Performance	Category rating and debriefing notes
Task Management	Planning & preparing			
	Prioritising			
	Providing & maintaining standards			
	Identifying & utilising resources			
Team Working	Co-ordinating activities with team			
	Exchanging information			
	Using authority & assertiveness			
	Assessing capabilities			
	Supporting others			
Situation Awareness	Gathering information			
	Recognising & understanding			
	Anticipating			
Decision Making	Identifying options			
	Balancing risks & selecting options			
	Re-evaluating			

***4** Good; **3** Acceptable; **2** Marginal; **1** Poor; **N** Not Observed

15

APPENDIX 15 OXFORD NOTECHS SCORE SHEET

NTS category		Surgeon team	Anaesthetic team	Nursing team
Leadership & management	Leadership Maintenance of standards Planning and preparation Workload management Authority & Assertiveness			
Teamwork & cooperation	Team building/maintaining Support of others Understanding team needs Conflict solving			
Problem solving & decision making	Definition & diagnosis Option generation Risk assessment Outcome review			
Situation awareness	Notice Understand Think ahead			

1 consistent	2 inconsistent	3 consistent	4 inconsistent	5 inconsistent	6 consistent	7 inconsistent	8 consistent
Behaviour compromises patient safety and effective teamwork		Behaviour in other conditions could directly compromise patient safety and effective teamwork		Behaviour maintains an effective level of patient safety and teamwork		Behaviour enhances patient safety and teamwork, a model for all other teams	

Range: 12-96

APPENDIX 16 OSCAR SCORE SHEET

0 = Team Severely Compromised 1 = Team Compromised
2 = Slight detriment to team 3 = Team neither enhanced or hindered
4 = Moderate enhancement to team 5 = High level of enhancement to team
6 = Highly effective in enhancing teamwork

COMMUNICATION

Anaesthetic Group (A)	**Individual Behaviour Ratings**							**Overall Imp (0-6)**
Informs team whether patient is making respiratory effort	0	1	2	3	4	5	6	
Informs team of any other relevant clinical signs e.g. dilated pupil, obvious injuries, signs of aspiration	0	1	2	3	4	5	6	
Communication to team that they plan to intubate the patient if required	0	1	2	3	4	5	6	
Requests patient history on arrival and communicates details to team, if required	0	1	2	3	4	5	6	
Physician Group (P)								
Reviews patient history and notes and communicates relevant details clearly to the team	0	1	2	3	4	5	6	
Clear instructions communicated to the team regarding the arrest protocol	0	1	2	3	4	5	6	
Encourages communication from sub-teams, and encourages team members to give opinions	0	1	2	3	4	5	6	
Nurse Group (N)								
Provides clear information about arrest events on arrival of arrest team	0	1	2	3	4	5	6	
Senior nurse provides clear, audible requests to junior nurse when requesting equipment e.g. additional iv bags	0	1	2	3	4	5	6	
Instructs other nurses on ward clearly how to assist with arrest or other ward duties as appropriate	0	1	2	3	4	5	6	

CO-OPERATION

Anaesthetic Group (A)	**Individual Behaviour Ratings**							**Overall Imp (0-6)**
A-group provides information on request from M-group (e.g. about the airway)	0	1	2	3	4	5	6	
A-group assists M-group in decision making in difficult scenarios	0	1	2	3	4	5	6	
Physician Group (P)								
Responds to questions from other team members about decisions made regarding the arrest	0	1	2	3	4	5	6	
Supports less experienced members of M-group, and compensates for their lack of experience	0	1	2	3	4	5	6	
Nurse Group (N)								
Provide support and assistance to A-group and M-group when needed e.g. finding airway adjuncts	0	1	2	3	4	5	6	
Help M-group locate items not routinely stocked on trolley, or missing from the trolley	0	1	2	3	4	5	6	
Assist M-group with extra tasks e.g. sending bloods, contacting family, contacting labs etc	0	1	2	3	4	5	6	

CO-ORDINATION

Anaesthetic Group (A)	**Individual Behaviour Ratings**							**Overall Imp (0-6)**
Information provided about changes in patient condition as they occur	0	1	2	3	4	5	6	
A-group co-ordinate team to move patient e.g. floor to bed, up bed	0	1	2	3	4	5	6	
Physician Group (P)								
Notifies N and A groups of anticipated further requirements for patient resuscitation	0	1	2	3	4	5	6	
Within M group, co-ordinates tasks such as taking of bloods, sending samples, sending ABG etc	0	1	2	3	4	5	6	
Nurse Group (N)								
Prepare Resus Trolley for use by team by bringing to bedside, turning monitor on etc	0	1	2	3	4	5	6	
Prepare further drugs in readiness for their next required use e.g. prepare next adrenaline minijet	0	1	2	3	4	5	6	
A Senior Nurse (Sister) is always present to provide backup to Staff Nurse	0	1	2	3	4	5	6	

LEADERSHIP

Anaesthetic Group (A)	**Individual Behaviour Ratings**							**Overall Imp (0-6)**
Advises team on best management, and contingency plans for patient, and takes lead if required	0	1	2	3	4	5	6	
Anaesthetist assertively takes a lead in Airway control and Ventilation on arrival at arrest	0	1	2	3	4	5	6	
Lead Anaesthetist supervises and supports staff lacking familiarity with tasks or equipment	0	1	2	3	4	5	6	
Physician Group (P)								
Takes a lead and clearly instructs assistants with requirements for arrest and/or defers leadership as required if appropriate	0	1	2	3	4	5	6	
Supervision given to staff lacking experience or familiarity with tasks or equipment	0	1	2	3	4	5	6	
Instructs N-group of additional requirements e.g. recent blood results from computer, to call the family	0	1	2	3	4	5	6	
Nurse Group (N)								
Takes a lead with initial Basic Life Support attempts until Arrest Team arrive	0	1	2	3	4	5	6	
Supervision and support given to junior or inexperienced members of N-team	0	1	2	3	4	5	6	

MONITORING

Anaesthetic Group (A)	**Individual Behaviour Ratings**							**Overall Imp (0-6)**
Maintains monitoring of patient condition, signs of respiration, other clinical signs	0	1	2	3	4	5	6	
Checks ventilation is adequate with regular blood gas analysis and amends ventilation accordingly	0	1	2	3	4	5	6	
Confirms drug identity by checking syringe labelling prior to drug administration	0	1	2	3	4	5	6	
Physician Group (P)								
Maintains awareness of activities of other teams e.g. anaesthetist intubating	0	1	2	3	4	5	6	
Monitors progress of resuscitation protocol with careful checking of time, and constant reassessment of limb of protocol and "extra considerations"	0	1	2	3	4	5	6	
Checks team condition e.g. monitors for fatigue in team members from CPR and suggests team members change roles, take turns etc	0	1	2	3	4	5	6	
Nurse Group (N)								
Monitors patient dignity and considers well-being of other patients nearby	0	1	2	3	4	5	6	
Maintains awareness of the needs of M and A groups	0	1	2	3	4	5	6	

DECISION MAKING

Anaesthetic Group (A)	**Individual Behaviour Ratings**							**Overall Imp (0-6)**
Prompt identification of the problem	0	1	2	3	4	5	6	
Rapidly and clearly outlines a strategy or plan, and asks for equipment	0	1	2	3	4	5	6	
Anticipates potential problems and prepares accordingly – e.g. asks for further blood crossmatched	0	1	2	3	4	5	6	
Physician Group (P)								
Rapidly decides an appropriate course of action for continued resuscitation	0	1	2	3	4	5	6	
Uses the team as a whole to help develop options – asks for opinions and processes them decisively	0	1	2	3	4	5	6	
Nurse Group (N)								
Prompt decision making during initial resuscitation attempts	0	1	2	3	4	5	6	
Anticipates potential problems A and M teams may encounter e.g. pulls bed out from wall, clears area etc	0	1	2	3	4	5	6	
Appropriate decision making regarding timing of initial decision to put out a cardiac arrest call	0	1	2	3	4	5	6	

APPENDIX 17 OTAS SCORE SHEET

EXEMPLAR BEHAVIORS

Anesthetists (A)	Nurses (N)	Surgeons (S)
Pre-operative Stage		
• Updates theater manager on any changes to case list	• Scrub nurse mediates progress of case through proactive communication	• Changes in the operation or case list communicated to all concerned
• Confirms patient details and condition with patient and informs N	• Confirms patient-specific requirements with A and S	• S talks to team and encourages communication from subteams
• Verbal communication to theater team on patient transfer and setup	• Communicate any problems regarding setup, provisions and staffing to team	• Verbal confirmation of procedure and intraop requirements
Intra-operative Stage		
• Asks surgeons if patient positioning is OK	• (SN) repeats surgeon's requests, confirming requirements	• Asks team if all are prepared to begin the operation
• Provides update on patient condition and anything administered to patient		• Asks A if ready to start the operation
• A inquires about operation and patient progress	• SN provides clear and audible requests for provisions to (CN)	• Requests and instructions to team communicated clearly and effectively
	• Swabs, needles, and instruments count confirmed verbally between CN and SN	• Provides information to whole team on progress
		• S informs the team of technical difficulties and/or changes of plan
		• S informs A of bleeding
Post-operative Stage		
• A instructs team on patient transfer to trolley	• Provides information concerning surgical procedure and patient condition to recovery nurses	• Informs and instructs team on any new patient requirements
• Asks team if ready to transfer patient and instructs on process		• Comments on work done in this case
• Information on patient condition and drugs provided to recovery nurse	• Recovery nurse confirms information transferred from theater team	
• A informs S about special needs for analgesia	• Ensures that patient documents are with patient in recovery	

SUMMARY SCALE

6 The team exchanged information proactively and politely. Case-specific communication was clearly audible and well articulated. The team made a concerted and consistent effort to maintain open communication to fulfile teamwork.

Team communication was highly effective in enhancing teamwork.

5 High level of enhancement to teamwork through communication.

4 Moderate enhancement to teamwork through communication.

3 Case-specific communication was acceptable, although members did sometimes seek clarification. The manner and effort of communication was reasonable. Team communication neither hindered nor enhanced team work.

Team communication neither enhanced nor hindered teamwork.

2 Slight detriment to teamwork through communication.

1 Teamwork compromised through poor communication.

0 The team did not communicate appropriately. Case-specific communication was unclear, and members consistently sought clarification and repeats or did not ask for clarification. The manner of communication was negative and unacceptable. This team had a problem communicating openly. Overall, the function of this team was hindered by poor communication.

Team communication severely hindered teamwork.

APPENDIX 18 SUMMARY OF VARIATION IN ORIGINAL METHOD OF TESTING, DATA COLLECTED AND STATISTICAL ANALYSIS FOR ANTS, OXFORD NOTECHS AND OSCAR

NTS assessment tool	Method of testing tool	Data assessed	Statistical tests used
ANTS	50 anaesthetists trained new to NTS rating, each rated 8, non-standardised videos of simulated anaesthetic scenarios	All elements (15) and category (4) scores, no global score	Accuracy of scores with percent agreement (± 1 scale point) and mean absolute difference IRR with r_{WG}
OTAS	150 live surgical procedures (general and urology) assessed in pairs (surgeon and psychologist)	Scores in five categories for three teams across three phases of surgery (pre-, intra- and post-operative), no global score	Differences in mean scores, IRR with Pearson's correlation coefficient for categories
Oxford NOTECHS	6 assessors (3 clinical, 3 human factors experts) assessed a total of 297 live surgical procedures in pairs (surgeon and human factors expert)	Total scores for categories in sub-teams and overall summated score (for all categories and all teams)	Differences in mean scores, IRR with ICC for global and category scores
OSCAR	2 clinical expert assessors, each rated 8, non-standardised videos of simulated cardiac arrest scenarios	Scores for each element, overall category score and a summated global score,	Descriptive statistics to explore mean scores between raters, IRR with ICC for global and category scores

APPENDIX 19 USABILITY EVALUATION QUESTIONNAIRE FOR ANTS, OXFORD NOTECHS AND OSCAR

General questions about the system (Questions were applied to all tools except where indicated)

1. Do you think the system was useful for structuring your observation of the film scenarios?

Yes / No If no, what was the problem?

2. Did it seem to address the key non-technical skill behaviours displayed by the individuals/team in the scenario?

Yes / No If no, what behaviours do you think were not addressed?

3. How easy was it to associate observed behaviours with the NTS tool's categories?

Very difficult / Difficult / Average / Easy / Very easy Please provide any specific comments about any or all of the categories:

4. Do you think there are any (non-technical) skills elements and/or categories missing from the list?

Yes / No If yes, what is missing?

5. Do you think there are any (non-technical) skills elements and/or categories in the list which are not necessary?

Yes / No If yes, which elements and/or categories are unnecessary?

6. Was the wording used for the category and element labels meaningful?

Yes / No If no, please describe where you thought there were problems

7. Were the descriptions for each category and element clear?

Yes / No If no, please describe which descriptions were unclear

8. Were the examples of 'good' behaviours helpful?

Yes / No Please give any comments (positive or negative) you may have

9. Were the examples of 'poor' behaviours helpful?

Yes / No Please give any comments (positive or negative) you may have

Questions about the rating scale

10. Please indicate how easy it was to use the rating scale provided:

Very difficult / Difficult / Average / Easy / Very easy

If you have any particular concerns please explain

11. Do you think the rating scale gave you enough flexibility to rate the performance levels seen in the film clips?

Yes No

If no, would you have liked a longer or shorter scale?

longer / shorter

12. Did you use the comments section on the rating form?

Yes No

If yes, please say what sort of information you noted down e.g. explanation of the performance rating you gave

13. Did you have any problems with the design of the rating form?

Yes / No

If yes, please explain

14. Was the amount of background information you were given (ANTS excluded):

too much / just right / too little

15. Were the explanations of the different categories and behavioural markers adequate(ANTS excluded)?

Yes / No

If no, how do you think the explanations could be improved?

16. overall do you think you were able to use the NTS system effectively?

Yes / No

ANY OTHER COMMENTS?

APPENDIX 20 ANSWERS TO USABILITY QUESTIONNAIRE (QUESTIONS INCLUDED BELOW)

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ANTS																
HH	Y	Y	Easy	N	N	Y	Y	Y	Y	Very easy	Y	Y	N	N/A	N/A	Y
NC	Y	Y	Average	N	N	Y	Y	Y	Y	Easy	Y	N	N	N/A	N/A	Y
PG	Y	Y	Easy	N	N	Y	Y	Y	Y	Easy	Y	Y	N	N/A	N/A	Y
Oxford NOTECHS																
HH	Y	Y	Easy	N	N	Y	Y	Y	Y	Easy	Y	Y	Y	Just right	Y	Y
NC	Y	Y	Easy	N	N	Y	Y	Y	Y	Very easy	Y	N	N	Just right	Y	Y
PG	Y	Y	Easy	N	N	Y	Y	Y	Y	Easy	Y	N	Y	Too much	Y	Y
OSCAR																
HH	N	Y	Average	Y	N	Y	Y	N	N	Difficult	N	Y	Y	Just right	N	N
NC	N	Y	Difficult	Y	N	Y	Y	N	Y	Very difficult	Y	N	Y	Too little	N	N
PG	N	Y	Difficult	Y	N	N	Y	N	N	Difficult	N	N	Y	Just right	Y	N

1. Do you think the system was useful for structuring your observation of the film scenarios?
2. Did it seem to address the key non-technical skill behaviours displayed by the individuals/team in the scenario?
3. How easy was it to associate observed behaviours with the NTS tool's categories?
4. Do you think there are any (non-technical) skills elements and/or categories missing from the list?
5. Do you think there are any (non-technical) skills elements and/or categories in the list which are not necessary?
6. Was the wording used for the category and element labels meaningful?
7. Were the descriptions for each category and element clear?
8. Were the examples of 'good' behaviours helpful?
9. Were the examples of 'poor' behaviours helpful?
10. Please indicate how easy it was to use the rating scale provided:
11. Do you think the rating scale gave you enough flexibility to rate the performance levels seen in the film clips?
12. Did you use the comments section on the rating form?
13. Did you have any problems with the design of the rating form?
14. Was the amount of background information you were given (ANTS excluded):
15. Were the explanations of the different categories and behavioural markers adequate(ANTS excluded)?
16. overall do you think you were able to use the NTS system effectively?

APPENDIX 21 SUMMARY OF QUALITATIVE DATA FROM USABILITY QUESTIONNAIRE AND POST-STUDY MEETING

Comments about the systems overall		
ANTS	ADVANTAGES: The category and element descriptors are described in general terms making them easy to apply across all situations There is a comprehensive handbook with useful summary pages	DISADVANTAGES: Only anaesthetist in the team assessed Some behaviours cross categories leading to difficulty in determining where to ascribe a score It is recommended that faculty undertake a two-day training course in the use of the system
Oxford NOTECHS	ADVANTAGES: Assesses three sub-teams in theatre (surgeons, the anaesthetic team and the surgical scrub team) and provides scores for each of these sub-teams Provides a global team score NTS domains are limited to 4 increasing ease of use Structure of NTS domains similar to ANTS (making it a more readily usable than OSCAR)	DISADVANTAGES: Would require additional faculty with context specific expertise to use whole tool
OSCAR	ADVANTAGES: OSCAR allows a comprehensive assessment of NTS in three sub-teams during the management of arrest situations Examples of behaviours given are specific and comprehensive which may help when you are new to using the tool	DISADVANTAGES: Example behaviours quite prescriptive – if they don't happen an otherwise well performing team may be inappropriately marked down 6 domains are challenging to mark – might be helpful to combine e.g. co-operation and co-ordination Blurring of lines between roles of Anaesthetic and Medic group could make it difficult to score if not experienced in resuscitation Some example behaviours did not fit with the authors' experience of role responsibilities at an arrest e.g. overlap in decision making between Anaesthetic group and Medic group and clarity on who is leading in an arrest
Comments about the rating scale		
ANTS	ADVANTAGES: Provides option of marking a behaviour "not observed" Scoring system on one page with room for notes	DISADVANTAGES: Only scoring anaesthetic NTS Problem with ceiling effect using 1-4 scale Does not provide a global rating
Oxford NOTECHS	ADVANTAGES: Scoring system all on one page Less risk of ceiling effect with 1-8 scale Descriptors provided for each domain	DISADVANTAGES: Not enough space to make notes on score sheet Suggested that a starting point of 6 for each assessment but automatically biases to higher end of scale - points 1-5 describe sub-optimal behaviour The descriptors "consistent" and "inconsistent" were challenging to use in some scenarios No option to mark a behaviour as "not observed"
OSCAR	ADVANTAGES: Score sheet is comprehensive with descriptors for each behaviour under consideration Less risk of ceiling effect with 1-6 scale	DISADVANTAGES: Not much distinction between a score of 5 (high level of enhancement to team) and a score of 6 (highly effective in enhancing teamwork) No option to mark a behaviour as "not observed" Score sheet covers 3 pages making it challenging to move back and forward between domains during marking period Scoring system covers three professional groups - would necessitate additional faculty More space to make notes would be good

APPENDIX 22 SASI STUDY: FACULTY GUIDELINE, PHYSIOLOGICAL PARAMETERS AND EXPECTED CANDIDATE ACTIONS FOR SCENARIO A

SASi Scenario A: Faculty Guideline

Main learning outcomes:

1) Improved understanding of the importance of good SA and communication in rapidly changing situations
2) Improved ability to work in an effective and efficient way as a team
3) Improved ability to recognise and manage fast AF in ICU patient
4) Greater confidence in the use of DC cardioversion for AF
5) Greater confidence in ALS management of asystolic cardiac arrest

Case Summary (use as pre-scenario briefing for candidates)

You have been asked to take over the care of a 75 year old man - Winston Moore, MRN Y1234567
Day 3 Adult ICU. Working diagnosis: Urosepsis.
PMHx: Hypertension
DHx: Atenolol; Aspirin; Simvastatin. NKDA
SHx: Non-smoker.; Minimal alcohol intake; walks with stick; lives in bungalow with wife; independent
Reasons for ICU admission:
- Intubated for low GCS
- Vasopressor BP support
- Monitoring of Acute Kidney Injury and fluid management

Current Issue: Not waking up after sedation has been off for 14 hours

History:
Patient presented to the ED with 2 days worsening confusion and reduced GCS – wife found him collapsed in bathroom. On antibiotics (co-amoxiclav) for a UTI (Day 4 of a course). In ED was hypotensive, tachycardic, pyrexial and drowsy. Intubated due to reduced GCS despite IV fluid resuscitation. CT Head scan normal on admission. Treated for urosepsis with IV fluids and antibiotics. Transferred to ICU after CT Scan.

Management since admission to ICU (3 days):
Ventilated via ET Tube; Noradrenaline for BP support with ECHO guided fluid therapy. Sedated with fentanyl and propofol. IV antibiotics (co-amoxiclav and SSG). Sedation stopped 14 hours ago as patient improving – not waking up. NG feed has been off since 0200
Planned for CT head scan to rule out intra-cranial pathology this evening. Booked for CT in 2 hours.

Scenario summary:

Approximately 10pm on AICU. Bedside nurse of this patient is asked to take an emergency admission to AICU – needs to hand over her current patient to the new nurse. Nurse 2 (study participant) receives handover and is advised to continue as normal.
During assessment short runs of AF begin and then becomes persistent, but no fall in BP.
Patient will not respond to pharmacological measures (e.g. electrolyte replacement, amiodarone) to treat AF and will become haemodynamically unstable, requiring cardioversion. Clinical team should proceed with DC Cardioversion: Brief period of sinus rhythm after 1st shock but will revert to Fast AF. After 3rd shock patient goes into asystolic cardiac arrest. Clinical team should manage Cardiac Arrest according to ALS protocol. ROSC after 1 cycle CPR. Scenario ends following ROSC.

Clinical situation	SimMan™ Settings	Expected candidate actions
PHASE 1 Bedside nurse (1) hands over the patient to her colleague (Nurse 2) as she is going to take a new admission – patient is due a set of observations and ABG. Nurse 2 takes over care Short runs of AF appear on patient monitor – progressing into persistent AF – no haemodynamic compromise Doctor joins scenario when requested to manage Fast AF Patient unconscious (off sedation) ; Intubated and ventilated; low dose noradrenaline running No response to IV amiodarone infusion No response to IV magnesium infusion Transient drop in HR to 145 with fluid bolus	**A** COETT Size 8.0 **B** Ventilator – BILEBEL 15/5, FiO2 0.30, TV 540; RR 20; SpO2 98%; EtCO2 4.8; clear chest **C** HR 95 sinus; IABP 100/60 (73); CRT 3 secs; warm peripheries; Arterial line 20G Rt radial; RIJ CVP line 4 lumen; 2 cannulas – left (22G) hand and right (18G) hand Noradrenaline running 0.08mcg/kg/min **D** GCS E1 VT M4; Pupils equal and reactive 4mm Fentanyl/propofol attached (not running) **E** NGT in-situ – feed currently disconnected Short runs AF rate 135bpm gradually appear – evolves to Fast AF rate 160bpm; BP drops slightly to 95/58 (70) HR drops transiently to 145 bpm after fluid bolus - not sustained ABG	• **Nurse 2** - receives handover from Nurse 1 - performs ABCDE assessment and records set of observations - notices runs of AF - runs ABG – should observe low K+ (3.2); recognises need to top up (no K+ prescribed) - performs 12 lead ECG - notices pyrexia - gives paracetamol (on PRN) for fever - calls for doctor to review patient - gives IV magnesium at doctor's request • **Doctor 1** - Prescribes K+ infusion - Requests septic screen (not done since admission) including peripheral blood cultures, CVP line culture, urine culture, BAL, CXR - Checks lab blood results – notices low Magnesium 0.68 – requests for this to be replaced. - Reviews 12 lead ECG – Fast AF – no ischaemia - Suggests and prescribes IV fluid bolus +/- IV magnesium+/- IV amiodarone - Reviews existing microbiology results – 'no growth' - Consider new antibiotic - Consider gentamicin level and repeated dose IV fluid bolus given IV potassium started IV Magnesium (Dose 2g) started

NB: TEAM MIGHT DECIDE TO ELECTRICALLY CARDIOVERT RATHER THAN GIVING DRUGS FIRST. IF SO, MOVE ON TO APPROPRIATE SECTION IN STAGE 2 POST CARDIOVERSION**	12 lead ECG CXR – can request (does not happen); can look at admission CXR Temp if checked is 37.2°C	IV amiodarone started – 300mg over 1 hour (can be reduced to 20-30 mins) then 900mg over 23 hrs (probably avoid IV beta blocker/calcium channel blocker in view of low dose noradrenaline requirement)
PHASE 2 = 60 SEC AFTER START OF RATE CONTROL DRUG Patient develops haemodynamic instability Back up nurse and doctor arrive (when called) Clinical team attempt DC cardioversion** If sedation given BP drops further – improves with noradrenaline After 1st shock patient reverts temporarily into sinus rhythm (100 bpm) but only for 20 seconds – then resumes Fast AF (rate 175bpm) After 2nd shock no effect	B SpO2 94%; EtCO2 3.9 C HR 160bpm AF; IABP drops gradually to 78/45; clammy; SpO2 increases to 99% if FiO2 increased IABP increases to 86/48 (60) with increase in noradrenaline/fluid bolus If sedation given– IABP drops to 75/40 1st shock – converts to sinus rhythm 100bpm (IABP 90/45) – for 20 seconds – then reverts to AF 170bpm 2nd shock – no change 3rd shock – asystole – immediate loss of ECG trace and loss of IABP trace	Recognition of haemodynamic instability Call for help – back up nurse and back up doctor arrive Recognition of need for DC Cardioversion Consideration of sedation e.g. IV propofol/midazolam small dose Increase FiO2 to 100% Safe operation of defib - paddles attached - monitoring leads attached - Synchronises defib - Charge to 150J - Checks no oxygen supply open - Checks no people near patient - Delivers shock

After 3rd shock patient becomes asystolic – loss of IABP trace on monitor		Notices conversion to Sinus Rhythm – starts to think about 12 lead ECG Notices reversion to Fast AF Performs 2nd DC shock (200J; as above) Performs 3rd DC shock (200J; as above)
PHASE 3 = AFTER 3RD SHOCK Cardiac Arrest (asystole) following 3rd DC shock ROSC shortly after 1st dose adrenaline given (team should continue CPR for 2 minutes in spite of this) ECG trace and Cardiac output detected on pulse check and arterial line after 2 minutes CPR Patient hypotensive but in sinus rhythm Scenario ends	B No SpO2 trace; No EtCO2 trace; Ventilator alarms (apnoea) and switches to BIPAP 20/5; rate 16; FiO2 as per team control C Asystole; IABP no output Defib monitor – shows CPR amplitude and output ECG and IABP trace show ROSC shortly after adrenaline Following 1 cycle CPR and 1mg adrenaline: B SpO2 96%; EtCO2 3.4 C HR 110 (sinus); IABP 100/60	Confirm Cardiac Arrest Start CPR – 100/min continuous chest compressions; mandatory ventilation Give 1mg adrenaline ASAP Continue CPR for 2 minutes Consider reversible causes of Cardiac Arrest – Hypoxia; hypovolaemia; Electrolyte disorders; Hypoglycaemia; Hypothermia; Thrombosis; Tension Pneumothorax; Tamponade; Toxins Re-check rhythm after 2 minutes

SASi Scenario B: Faculty Guideline

Main learning objectives:

1) Improved understanding of the importance of good SA and communication in rapidly changing situations

2) Improved ability to work in an effective and efficient way as a team

3) Improved ability to recognise and medical management of bronchospasm

4) Improved confidence in recognising and managing tension pneumothorax and VT (Torsades de Pointes)

5) Improved confidence in recognising and managing of VF arrest

Case Summary (use for pre-scenario briefing)

- 36yo man – Henry Cooper - MRN X1234567
- D1 AICU. Working Diagnosis: Polytrauma – splenic laceration, small bowel perforation and right humeral shaft fracture
- PMHx: Childhood asthma (not on regular inhalers)
- DHx: nil, NKDA
- SHx: 'Social' smoker; 20 units/week alcohol; works full time as a vet

Reason for ICU admission:

- Intubated following trauma laparotomy
- Low dose vasopressor for BP support
- Major haemorrhage intra-operatively and large transfusion
- Localised peritoneal contamination (small bowel perforation)
- Open abdomen – concern regarding intra-abdominal pressure

Current issue:

Just arrived in ICU from theatre

History:

Patient BIBA at 4pm today. Was out riding his horse on country lane when horse bucked at a car. Patient thrown from horse who then kicked his abdomen. Passing dog walker called 999. No LOC. GCS 15 throughout. In ED complained of severe abdominal pain and right shoulder pain. He was hypotensive, tachycardic, with bruising to abdomen and bruising and swelling to right upper arm. Lactate 4 and Hb 13 on initial blood gas. No respiratory or neurological concerns. CT scan revealed splenic laceration, extensive blood in abdomen and small bowel perforation. X-ray Right arm – right humeral shaft fracture (non displaced). Cervical spine cleared radiologically and CT chest and head normal.

Management prior to ICU admission:

Trauma Call in ED. Major haemorrhage protocol activated. Tranexamic Acid 1 gram given in ED and infusion just completed in theatre. Went from ED to CT then straight to emergency theatre. Underwent midline laparotomy; findings of splenic rupture and small bowel perforation; procedure - splenectomy, small bowel resection and anastomosis. Abdomen left open. Approximately 3 litres EBL. Received 2 litres Hartmann's, 8 units RBC, 4 FFP, 2 platelets given in total. General anaesthetic – RSI (ketamine, fentanyl, rocuronium); grade 1 intubation; Size 7.0 COETT tied at 22cm at teeth. RIJ CVP. Rt Radial arterial line. 1 x 14G cannula left ACF; 1 x 16G cannula right ACF; 1 x 18G cannula left hand. Urinary catheter. Paracetamol given at 21:30. Augmentin and Metronidazole given towards end of laparotomy at 21:30. Total 400mcg fentanyl given in theatre. Anaesthetist observed multiple ventricular ectopics intra-operatively (no treatment required). Patient transferred to ICU intubated.

Scenario summary:

Approximately 10pm on AICU. Bedside nurses and doctor receive handover from theatre anaesthetist (as above). Baseline ECG (pre-operative) shows long QT – given to team. Whilst assessing patient, a gradual deterioration occurs

with hypoxia, signs of bronchospasm, increased airway pressure and decreased tidal and minute ventilation over 2-3 mins. No other features of anaphylaxis and no evidence of pneumothorax. Vital signs and ventilator parameters initially improve with bronchospasm treatment (nebulisers and steroids). There is subsequently a further rapid deterioration with marked respiratory and then cardiovascular compromise – airway pressures rise significantly, patient desaturates, then runs of VT and hypotension. Transitions into continuous polymorphic VT (Torsades) with output initially. Examination indicates left sided tension pneumothorax. Shortly after decompression of pneumothorax, monitor shows VF. Conversion to sinus rhythm after 1st DC shock – team should resume 2 minutes CPR following shock according to ALS algorithm prior to rhythm check. After 2 mins, cessation of CPR allows confirmation of ROSC, sinus rhythm with sustained BP. Scenario ends following confirmation of ROSC. Note baseline ECG shows long QT.

Clinical situation	Sim Man Settings	Expected candidate actions
PHASE 1 Bedside nurse and junior registrar on call take handover from theatre anaesthetist	**A** COETT Size 7.0 **B** Ventilator – BILEVEL 18/5, FiO2 0.50, TV 500; RR 16; PAP 22; SpO2 98%; EtCO2 4.6; crepitations at lung bases and expiratory wheeze throughout **C** Cool peripheries, looks pale, cap refill 2 seconds. HR 105 bpm. Sinus rhythm. IABP 115/85. IV cannula 14G Lt ACF, 18G DLH, 16G Rt ACF. Rt radial arterial line. RIJ CVP line. Noradrenaline 0.15 mcg/kg/min. IV Hartmann's 100ml/hr. TEDS and flowtrons. Baseline ECG shows Long QT. D Sedated – propofol 200mg/hr and fentanyl 150mcg/hr. Pupils equal 1mm and reactive. RASS – 4. E Temp 37.7. NG Tube (Ryles) in situ on drainage. Open abdomen (dressed); abdominal drains LUQ	- Handover received from anaesthetist - ABCDE assessment of patient and recording of observations - Request and review post-op ECG, ABG and CXR

	and LIF (haemo-serous fluid); urinary catheter in situ (urometer). Bruising/swelling around right upper arm. BM 4.6.	
	Transition over 2-3 minutes (depending on candidate actions) to…	
Patient deteriorates clinically over 2 - 3 minutes (depending on candidate actions) due to bronchospasm. Becomes increasingly hypercapnic and difficult to ventilate with an obstructive capnography trace, rising airway pressures and falling minute ventilation.	A ETT remains patent; low volume clear secretions on suction B Ventilator alarming; ETCO2 6.2; Obstructive capnograph trace; TV 280; dyssynchronous ventilation; SpO2 85% (increases to 94% on 100% FiO2) C HR 130; IABP 140/90	- Detect wheeze - Detect rise in ETCO2 - Acknowledge ventilator alarm - Detect increase in airway pressure, change in shape of capnograph (obstructive), reduction in tidal volume - Diagnose bronchospasm - Look for rash - Suction ETT - Check most recent ABG - Acknowledge childhood asthma - Consider differential diagnosis of respiratory deterioration including ○ Asthma ○ Transfusion reaction ○ Anaphylaxis ○ Fluid Overload - Repeat ABG - Request CXR - Request ECG
Patient improves in response to treatment with nebulisers and steroids.	B SpO2 98%; TV 440; Less obstructive capnograph trace; ETCO2 5.0 C HR 120 sinus tachycardia; IABP 130/85	- Administer nebulised salbutamol and ipratropium bromide - Adjust ventilator settings: ○ Increase FiO2 1.0 ○ Reduce I:E ratio (prolong expiratory time) - Consider administration of ○ Magnesium Sulphate IV 2g over 20 minutes ○ Hydrocortisone 100mg ○ Frusemide 20mg - Repeat ABCDE assessment

PHASE 2 = 3 MINS AFTER MEDICAL TREATMENT STARTED Onset of tension pneumothorax Rapid deterioration in oxygenation and ventilation with associated hypotension and tachycardia Short runs of VT followed by persistent polymorphic VT Improvement in respiratory and cardiovascular function following needle decompression	**Transition over 60 seconds to:** B SpO2 70%; TV 200ml; ventilator alarming high pressures; obstructed capnography trace; ETCO2 2.1; wheeze on left side; no breath sounds heard on right side; hyper-resonant on right side C HR 120 sinus tachycardia; BP 70/40 (sustained); Short runs of VT rate 140 (30 secs) followed by permanent polymorphic VT **Post needle decompression:** B SpO2 94%; wheeze bilaterally; percussion note equal; breath sounds present bilaterally (harsh on left side); TV 420 C HR remains 140 VT (Torsades); BP 100/60; cold and clammy	- Observe hypoxia - Observe hypercapnia - Observe deterioration in ventilatory parameters - Observe runs of polymorphic VT - Observe hypotension - Observe fall in ETCO2 - Reassess ABCDE - Detect lack of breath sounds on right side - Detect hyper-resonant percussion note on right side - Recognise potential tension pneumothorax and emergency treatment required - Needle decompression of right sided pneumothorax – 2nd IC space; mid clavicular line - Call for more senior help - Anticipate need for surgical chest drain - Consider need for cardioversion - Consider treatment with IV Magnesium 2g
PHASE 3 = 3 MINS AFTER NEEDLE DECOMPRESSION (OR AFTER DCCV) Cardiac Arrest – VF	B Loss of SpO2 and EtCO2 trace; Ventilator continues as set	Observe - rhythm change on ECG (VF) - absence of IABP/SpO2 trace - loss of ETCO2 Check pulse and confirm Cardiac Arrest

ROSC after 1st DC shock (team should continue CPR for 2 minutes in spite of this) Following 2 mins CPR, ROSC confirmed with sustained BP, SpO2 and ETCO2 trace. Scenario ends following confirmation of ROSC	C ECG and defib monitor reveal VF; loss of IABP; no pulses **Following 1 DC shock, whilst CPR ongoing:** B SpO2 poor trace; EtCO2 interrupted pattern (B SpO2 96%; EtCO2 C HR 110 (sinus); IABP 100/60	Start CPR – 100/min continuous chest compressions; mandatory ventilation Connect defibrillator pads Confirm rhythm VF - Recognise shockable rhythm Perform safe defibrillation ASAP - pads attached - monitoring leads attached - Charge to 150J - Chest compressions continue during charging - Checks no oxygen supply open - Checks for safety (O2 and staff away) - Delivers shock - Should NOT give adrenaline (would be due with amiodarone after 3rd shock) Resume CPR immediately for 2 minutes Consider reversible causes of Cardiac Arrest – Hypoxia; hypovolaemia; Electrolyte disorders; Hypoglycaemia; Hypothermia; Thrombosis; Tension Pneumothorax; Tamponade; Toxins Consider review of previous 12 lead ECG in view of Torsades de pointes – reveals Long QT Consider IV Magnesium 2g IV Re-check rhythm after 2 minutes - sinus rhythm confirmed - pulse detected - confirms ROSC - Initiates post Cardiac Arrest management

APPENDIX 24 SASI STUDY: GDTA AND SAGAT QUESTIONS FOR SCENARIO A

AICU SASi study SAGAT Scenario A

MAJOR GOAL - SUCCESSFUL MANAGEMENT OF DETERIORATING PATIENT IN FAST AF

Sub-goal 1 – recognition and initial management of rhythm change from to SR to AF (pause at point where initial treatments have been planned or begun):

Decisions/actions:

Recognise and diagnose problem

Decide necessary investigations

Consider treatment options and choose

Project which other treatments may be necessary

Consider need for additional support

SA requirements:

Colour key:

Level 1 SA (perception)

Level 2 SA (comprehension)

Level 3 SA (projection)

Level 1 (perception)

Evidence of comorbidity and relevant PMH

Data from monitor: ECG (rhythm and rate), IABP, SpO2

Data from blood tests

Check temperature

Level 2 (comprehension)

Correctly diagnose AF, understands need to check BP and SpO2 repeatedly

Understands need to check electrolytes, including rechecking K

Uses holistic approach: considers intravascular volume, CVP line placement, temp

Level 3 (projection)

Anticipates possibility of cardiovascular instability – low BP, increased HR, fall in O2 delivery

Anticipates need to replace electrolytes (K and Mg)

Anticipates need to administer amiodarone, IV fluids

Anticipates need to investigate sepsis

Anticipates requirement for additional assistance

SAGAT Questions:

1 - What was the patient's heart rate at the start of the scenario (you may not have been in the room)?

2 - Has the rate gone up or down or is it unchanged?

3 - Do you anticipate a further change in the HR?

4 - What was the rhythm at the start of the scenario?

5 - What is the rhythm now?

6 - Was there any prior warning?

7 - What was the BP at the start of the scenario?

8 - What is the BP now?

9 - Are you concerned about the blood pressure?

10 - What was the last potassium level?

11 - Are you concerned about the potassium level?

12 - What are the next steps?

Sub-goal 2 – recognition of deterioration in cardiovascular stability (pause at point of definitive decision to cardiovert or just prior to DC cardioversion):

Decisions/actions:

Observe and reassess regularly

Recognition of signs of deterioration

Holistic approach to diagnosis in light of changes

Recognition of increasing team activity

Consider need for additional assistance and prepare

Prepare for potential problems (including calling for necessary equipment and acquisition of equipment/drugs etc)

SA requirements:

Level 1 (perception)

Data from monitors: ECG, IABP, SpO2, ETCO2

Monitor progress with tasks

Level 2 (comprehension)

Understands that patient is now deteriorating and has not responded to initial management

Understands that DC cardioversion should be the next step

Projection (level 3)

Anticipates needs for senior assistance as appropriate

Anticipates treatment failure or further deterioration

SAGAT questions:

13 - What is the patient's HR?

14 - What is the patient's BP?

15 - What is your main concern at this point?

16 - Do you need additional resources? If yes, who?

17 - What are the next steps?

18 - What might go wrong?

Sub-goal 3 – Recognition of asystole and management (pause just after diagnosis of asystole):

Decisions/actions:

Regular reassessment of situation

Use appropriate ALS algorithm

Consider underlying causes

Consider additional assistance

SA requirements:

Level 1 (perception)

Data from monitor: ECG, IABP, SpO2, ETCO2

Pulse check

Level 2 (comprehension)

Recognise asystole

Use algorithm for asystole (including immediate delivery of adrenaline)

Understand possibility of failure to rescue

Projection (level 3)

Anticipates needs for assistance

Anticipates treatment failure

SAGAT Questions:

19 - What is the rhythm?

20 - Did the patient have a cardiac output?

22 - What are the next steps?

22 - -What might go wrong?

APPENDIX 25 SASI STUDY: GDTA AND SAGAT QUESTIONS FOR SCENARIO B

AICU SASi study SAGAT Scenario B

MAJOR GOAL - SUCCESSFUL MANAGEMENT OF SEVERE BRONCHOSPASM COMPLICATED BY TENSION PNEUMOTHORAX

Sub-goal 1 – recognition and initial management of bronchospasm and respiratory deterioration (pause at point where initial treatments have been planned or begun):

Decisions/actions:

Recognise and diagnose problem

Decide necessary investigations

Consider treatment options and choose

Project which other treatments may be necessary

Consider need for additional support

SA requirements:

Colour key:

Level 1 SA (perception)

Level 2 SA (comprehension)

Level 3 (projection)

Level 1 (perception)

Evidence of comorbidity and relevant PMH

Data from pre-operative ECG

Data from pre-operative blood tests

Data from patient: check tube level and listen to chest

Data from monitor: ECG, IABP, SpO2, ETCO2

Data from ventilator – specifically PAWP, VT, MV

Data from current blood tests especially ABG

Level 2 (comprehension)

Diagnoses bronchospasm, understands potential underlying causes and delivers treatment as appropriate, understands effects on airway pressure

Understands ventilation abnormalities and considers differential (for expiratory airflow obstruction)

Understands need to check ABG

Understands importance of CXR

Uses holistic approach: consideration of other causes of bronchospasm (e.g. looks for rash)

Level 3 (projection)

Anticipates need to alter ventilatory parameters or modes

Anticipates need to use additional bronchodilators

Anticipates possibility of pneumothorax

Anticipates other causes of change in CO2 trace and looks for them (e.g. circuit obstruction, ET tube misplacement)

Anticipates possibility of cardiovascular instability – low BP, increased HR, fall in oxygen saturations

Anticipates complications of long Q-T syndrome if this has been picked up (e.g. Torsades de Pointes)

Anticipates requirement for additional assistance

SAGAT questions

1 - What was the $ETCO_2$ at the start?

2 - What was the cause of the ventilator alarm at the beginning?

3 - What were the airway pressures when we paused?

4 - What was the $ETCO_2$ when we paused?

5 - Were the airway pressures within acceptable limits?

6 - Is the $ETCO_2$ trace normal?

7 - What is the most likely cause of the problem?

8 - What do you think will happen to the airway pressures?

9 - What do you think will happen to the SpO_2?

10 - What are the next steps?

Sub-goal 2 – recognition and management of tension pneumothorax and polymorphic VT (pause at point of needle decompression or decision to cardiovert):

Decisions/actions:

Observe and reassess regularly

Recognition of signs of deterioration

Holistic approach to diagnosis in light of changes

Recognition of increasing team activity

Consider need for additional assistance and prepare

Prepare for potential problems (including calling for necessary equipment and acquisition of equipment/drugs etc)

SA requirements:

Level 1 (perception)

Data from monitor: ECG, IABP, SpO2, ETCO2

Data from ventilator

Examination findings: Lack of breath sounds and hyper-resonant percussion note on the right side with wheeze on the left

Monitor progress with tasks

Level 2 (comprehension)

Understands patient has tension pneumothorax, is haemodynamically unstable and peri-arrest

Understands VT could be linked to long-QT

Understands emergency intervention (needle decompression of pneumothorax) is required

Understands need to consider necessary skill mix in team

Projection (level 3)

Anticipates needs for senior assistance as appropriate

Anticipates treatment failure or further deterioration

Anticipates requirement for chest drain

Anticipates requirement for pharmacological treatment of VT (e.g. Mg, amiodarone)

Anticipates requirement for DC cardioversion

SAGAT questions:

11 - What is the patient's HR?

12 - What is the patient's BP?

13 - What are the findings on respiratory examination?

14 - What did you notice about the ECG on the monitor?

15 - What is your main concern at this point?

16 - Do you need additional resources? If yes, who?

17 - What are the next steps?

18 - What might go wrong?

Sub-goal 3 – Recognition and management of Cardiac Arrest (VF) (pause at the point of diagnosis of Cardiac Arrest):

Decisions/actions:

Regular reassessment of situation

Use appropriate ALS algorithm

Consider underlying causes

Consider additional assistance

SA requirements:

Level 1 (perception)

Data from monitor: ECG, IABP, SpO2

Pulse check

Level 2 (comprehension)

Recognise shockable rhythm

Use algorithm for shockable rhythms (including delivering first shock as soon as possible)

Projection (level 3)

Anticipates requirement for immediate shock

Anticipates problems with potential conduction anomaly (reference to initial ECG)

Anticipates need for assistance

Anticipates treatment failure – communication with senior team

SAGAT questions:

19 - What is the rhythm?

20 - Did the patient have a cardiac output?

21 - What are the next steps?

22 - What might go wrong?

APPENDIX 26 SASI STUDY : OXSTAR CONSENT FORM AND CONFIDENTIALITY AGREEMENT

OxSTaR Centre

As a course delegate taking part in simulation sessions I understand the significance of confidentiality with respect to information concerning simulated patient scenarios. I also understand that confidentiality regarding fellow delegates must be maintained. I agree to report any violations of confidentiality that I become aware of to the Centre Coordinator.

I agree to adhere to the following guidelines:

- All information in the simulated scenarios and the debriefing sessions is privileged and confidential regardless of format: electronic, written, overheard or observed.
- I may view, use, disclose, or copy information only as it relates to the performance of my simulation training.
- The simulation lab is a learning environment all scenarios, regardless of their outcome, should be treated in a professional manner. At all times course participants should respect the learning needs of fellow delegates.
- The simulation mannequins are to be used with respect and be treated as if they were live patients. There are guidelines for delegates taking part in the simulation sessions; these will be highlighted in the introductory orientation session.
- All delegates will abide by any direction or instruction pertaining to safety given by a member of faculty or the OxSTaR Centre staff.
- All mobile phones should be silenced and delegates are asked to refrain from texting or emailing during the course.
- Food and drink are permitted in the seminar room but please do NOT take food and drink into the simulation room

In addition I agree:

- to be videoed for educational feedback purposes

Please tick one of the following:

I would like the recording destroyed after the session ☐

or

I consent to the video being kept for a maximum of 1 year (after which it will be destroyed) for:

a) Research ☐

b) Teaching purposes, which **may include the video being shown at meetings, training days or other teaching activities in the OxSTaR centre.** ☐

Do you have any physical conditions that may affect your participation in the scenarios? Y/N

Date: _____ Course: _____AICU SASI_____

Please indicate if you are a nurse or a doctor: NURSE/DOCTOR

How many years experience in AICU do you have?:

AICU SASi training

FEEDBACK QUESTIONNAIRE

1 How confident are you in your ability to work well in a team?

before training: 1 2 3 4 5 6 7 8 9 10

after training: 1 2 3 4 5 6 7 8 9 10

(Not very confident ..... very confident

2 How confident do you feel in your ability to recognise changes in a patient's condition?

before training: 1 2 3 4 5 6 7 8 9 10

after training: 1 2 3 4 5 6 7 8 9 10

(Not very confident .....very confident)

3 How confident are you in communicating safety critical information?

before training: 1 2 3 4 5 6 7 8 9 10

after training: 1 2 3 4 5 6 7 8 9 10

(Not very confident .....very confident)

4 How confident do you feel in your ability to recognise your limitations and know when to call for help?

before training: 1 2 3 4 5 6 7 8 9 10

after training: 1 2 3 4 5 6 7 8 9 10

(Not very confident .....very confident)

5 How disruptive to your learning were the pauses in the SAGAT scenario?

1 2 3 4 5 6 7 8 9 10

(Not at all disruptive .....very disruptive)

FREE TEXT QUESTIONS:

6. **What are the most useful things you are taking away from the teamwork training?**

7. **Did you achieve your learning outcomes?**

8. **Please comment on how you will use what you have learnt on the course today in your every day clinical practice.**

9. **Did you find the use of pauses in the SAGAT scenario helpful in your learning experience?**

10. **Would you recommend this training to others?**

11. **Any other comments?**

Thank you very much – your comments are used to change and improve our courses.

APPENDIX 28 SASI STUDY: SART QUESTIONNAIRE

Candidate ID: **Date of training:** **Scenario A / B**

Please answer these questions as objectively as you can – think back to the scenario you have just been in and make a judgment on how you felt and record it below

SITUATION AWARENESS	Description	Low						High
		1	2	3	4	5	6	7
1.DEMAND	Instability of situation							
	Variability of situation							
	Complexity of situation							
2.SUPPLY	Arousal							
	Spare mental capacity							
	Concentration							
	Division of attention							
3.UNDERSTANDING	Information quantity							
	Information quality							
	Familiarity							

Construct descriptions:

1. Rate the demand on your attention:

Instability of situation: how likely did you think it was that the situation would change suddenly?
Variability of the situation: describe the number of variables requiring your attention (was it high or low?)
Complexity of the situation: what was the degree of complication (the number of closely connected variables) of the situation?

2. Rate the supply of your own attentional resources:

Arousal: describe the degree to which you were ready for activity
Spare mental capacity: describe the amount of mental ability available to apply to new variables
Concentration: what was the degree to which your thoughts were brought to bear on the situation?
Division of attention: what was the degree to which your attention was divided in the situation?

3. Rate your understanding of the situation:

Information quantity: rate the amount of knowledge you received and understood
Information quality: rate the value of the knowledge communicated
Familiarity: rate the degree to which you were familiar with the situation

APPENDIX 29 NASA-TLX SCORE SHEET

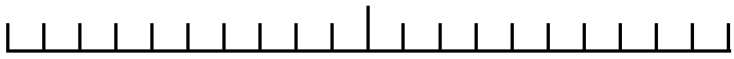
(electronic version available as an app:
https://humansystems.arc.nasa.gov/groups/TLX/tlxapp.php)

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

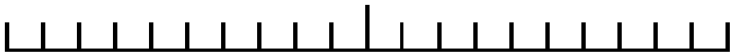
Name	Task	Date

Mental Demand — How mentally demanding was the task?



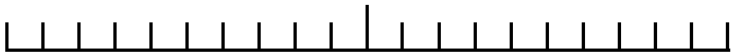
Very Low — Very High

Physical Demand — How physically demanding was the task?



Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?



Very Low — Very High

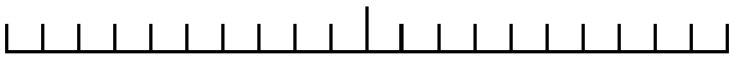
Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?



Very Low — Very High

APPENDIX 30 NASA TASK-LOAD INDEX, DESCRIPTION OF DOMAINS

NASA TASK-LOAD INDEX DESCRIPTION OF DOMAINS

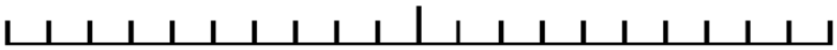
TITLE	ENDPOINTS	DESCRIPTIONS
MENTAL DEMAND	Low/High	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	Low/High	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	Low/High	How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
PERFORMANCE	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals?
EFFORT	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	Low/High	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

APPENDIX 31 SCORE SHEET FOR SASI GLOBAL PERFORMANCE SCORE

SASi Global Performance Score

Team No: Date:

Please rate global team performance on the following scale:



Extremely poor performance Excellent performance

Instructions for use

- Consider *technical* performance across the whole scenario
- Use GDTA for each scenario to consider level of *technical* performance in relation to task completion and decisions made – e.g. your assessment of tasks fully completed, timeliness of task completion, appropriateness of decisions

APPENDIX 32 RESULTS OF ANALYSIS OF CORRELATIONS BETWEEN SAGAT SCORES, EXPERIENCE AND WORKLOAD FOR PROFESSIONAL GROUPS

Correlations with overall SAGAT score	Scenario A		
	Doctors: Pearson's *r* (p-value)	Senior nurses: Spearman's ρ (p-value)	Junior nurses: Pearson's *r* (p-value)
Experience	-0.389 (0.34)	-0.094 (0.83)	-0.691 (0.06)
Workload (NASA TLX overall)	0.101 (0.81)	0.409 (0.32)	0.207 (0.62)
Workload (NASA TLX MD)	0.205 (0.63)	-0.553 (0.16)	-0.359 (0.38)
Workload (NASA TLX TD)	0.580 (0.13)	-0.064 (0.88)	-0.162 (0.70)
Workload (NASA TLX effort)	0.096 (0.82)	-0.270 (0.52)	0.038 (0.93)
Workload (NASA TLX frustration)	-0.272 (0.52)	0.587 (0.13)	0.339 (0.41)
Correlations with overall SAGAT score	**Scenario B**		
	Doctors: Spearman's ρ (p-value)	Senior nurses: Pearson's *r* (p-value)	Junior nurses: Pearson's *r* (p-value)
Experience	-0.314 (0.35)	0.188 (0.58)	0.078 (0.82)
Workload (NASA TLX overall)	0.287 (0.39)	0.229 (0.50)	0.269 (0.42)
Workload (NASA TLX MD)	0.409 (0.21)	-0.003 (0.99)	-0.467 (0.15)
Workload (NASA TLX TD)	0.466 (0.15)	0.098 (0.77)	0.080 (0.82)
Workload (NASA TLX effort)	0.408 (0.21)	0.345 (0.30)	-0.192 (0.57)
Workload (NASA TLX frustration)	-0.295 (0.38)	0.271 (0.42)	-0.319 (0.34)

APPENDIX 33 PUBLICATIONS AND PRESENTATIONS ARISING FROM THIS THESIS

Peer reviewed papers:

Higham H, Baxendale B. To Err is Human: Use of Simulation to Enhance Training and Patient Safety in Anaesthesia. *Br J Anaesth* 2017;**119**:i106–14. doi:10.1093/bja/aex302

Higham H, Greig P, Rutherford J, Vincent L, Young JD, Vincent C. Systematic Review of Validity, Reliability and Usability of Non-Technical Skills Assessment Tools for Use in Simulated or Real Clinical Environments in Healthcare. BMJ Quality and Safety (Submitted July 2018, accepted [with revisions] September 2018).

Presentations:

To Err is Human: Use of Simulation to Enhance Training and Patient Safety in Anaesthesia – invited lecture Royal College of Anaesthetists 25th Anniversary Meeting, London, March 2017

Simulation to Enhance Performance in Non-Technical Skills – invited key note Irish Association of Simulation meeting, Cork University, April 2017

Improving Safety in Non-Cardiac Surgery – invited lecture Royal College of Anaesthetists Cardiac Risk Symposium, London May 2017

Human Factors in Healthcare – invited key note - Oxford Deanery Anaesthetic Trainees Meeting, John Radcliffe Hospital, Oxford, January 2018

Human Factors and Team Training in Healthcare – invited key note Royal College of Surgeons of Ireland, Dublin, June 2018

Can we Pursue Excellence in Healthcare Through the Use of Simulation? – invited key note 10th Oxford Colloquium on Medical Education, St Anne's College Oxford, September 2018

Conference abstracts:

Higham H, Vincent L, Greig P, Warren R, Venes T, Chantler J, McKechnie S, Young D, Vincent C. *Use of the Situation Awareness Global Assessment Tool (SAGAT) and the Situation Awareness Rating Technique (SART) for Standardised Scenarios in Simulation Training for Intensive Care Teams*. Association for Simulated Practice in Healthcare Annual Conference, Southport, November 2018,

Higham H, Vincent L, Greig P, Warren R, Venes T, Chantler J, McKechnie S, Young D, Vincent C. *Analysis of validity and usability of the Situation Awareness Global Assessment Tool (SAGAT) during Simulation Training for Intensive Care Teams*. Intensive Care Society: State of the Art meeting, London, December 2018

Higham H, Greig P, Rutherford J, Vincent L , Young D, Vincent, C. *Systematic Review of Non-technical Skills Assessment Tools for use in Simulated or Real Clinical Environments in Healthcare*. International Meeting on Simulation in Healthcare, San Antonio, January 2019