OBJECTIVE MEASUREMENT OF NON-TECHNICAL SKILLS IN SURGERY

by

Jackie Soyoun Cha

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Industrial Engineering

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Denny Yu, Chair

School of Industrial Engineering

Dr. Steven J. Landry

School of Industrial Engineering

Dr. Robert W. Proctor

Department of Psychological Sciences

Dr. Dimitrios Stefanidis

Department of Surgery, Indiana University School of Medicine

Approved by:

Dr. Abhijit Deshmukh

To everyone who has supported me

ACKNOWLEDGMENTS

This work was supported, in part, by Intuitive Surgical Training and Education Research Grant.

This PhD journey took a village. As in the game of Mafia, there were rounds of mistaken judgement, second-guessing, and loss of villagers, but in the end, only the fun, laughter, and growth are remembered.

First and foremost, I would like to express my sincere gratitude to my advisor, and friend, Dr. Denny Yu. Thank you for believing in an undergrad who asked "what is human factors" many years ago and guiding me to become the researcher I am today. It has been a privilege to have been on this journey with you – we did it!

To my committee members, Dr. Dimitrios Stefanidis, Dr. Robert Proctor, and Dr. Steven Landry: thank you for your insights, support, and mentorship throughout the program. I will always be grateful for your contributions and investment to this work and to my development. I would especially like to thank the collaborators at Indiana University School of Medicine: Mr. Nicholas Anton, Dr. Dimitrios Athanasiadis, Dr. Sara Monfared, and Dr. Chandru Sundaram for their support in the research.

A special appreciation to the School of Industrial Engineering at Purdue University and to my labmates and colleagues in the Healthcare Ergonomics Analytics Lab (HEAL) and Intelligent Systems and Assistive Technologies (ISAT) lab. Many thanks to Dr. Abhijit Deshmukh, Dr. Brandon Pitts, Dr. Juan Wachs, Ms. Anita Park, and Ms. Leza Dellinger for their support and guidance throughout the program.

A heartfelt thanks to my friends for their encouragement and unforgettable memories: Ting Zhang, Glebys Gonzalez, Naveen Madapana, Nina Dutta, Yutaka Oshikiri, Shruthi Suresh, Benjamin Rachunok, and Gaojian Huang – you kept me sane.

Finally, thank you to my extended family but particularly to mom, dad, and Kenny for your constant nagging, support, and love. Yes, I am finally graduated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Non-technical skills (NTS) are cognitive and interpersonal skills that are relevant to task completion such as situation awareness, decision-making, teamwork, and leadership. NTS in clinical environments, such as surgery, have been identified to contribute to patient safety and team performance, which in turn affects clinical outcomes. Assessment tools of these skills in surgery exist; however, current evaluations are limited in that they require trained raters, are subjective, are time-intensive, and are checklist-based. Therefore, there is a need for objective measurement of NTS that addresses the limitations of the rating-based techniques. The purpose of this Ph.D. dissertation work is to identify physiological and behavioral metrics that measure NTS objectively and investigate the application of objective metrics to measure intraoperative NTS of surgeons. Through a scoping review of engineering, behavioral science, and medical literature, behavioral and physiological metrics that quantified NTS constructs of surgeons were identified. The synthesized literature was used to build a framework integrating objective metrics to NTS constructs. To develop an objective model of surgeons' NTS, subjective and objective behavioral data of surgeons were collected in the operating room and prediction models were created. Results found that objective metrics such as communication, speech, and proximity features can be used to predict subjective NTS. Furthermore, objective task features (e.g., time and number of incidents during an operation) has the potential to also model subjective NTS, and these task features can be predicted by the behavioral metrics; thus, triangulation is obtained with the three NTS metrics: subjective score, objective behavioral metrics, and task performance metrics. The relationship between the two objective metrics shows the possibility of achieving a fully objective model of surgeons' NTS. The consolidation of current objective measurement techniques can provide a foundation in further understanding NTS beyond assessments based on observed behaviors, and the developed models can be expanded and implemented for real-time NTS assessment of clinical teams to improve patient care.

1. INTRODUCTION

The impact of surgical team members' non-technical skills (NTS) to technical performance and patient safety has been gaining increasing attention (Gawande et al., 2003; Hull et al., 2012; Leuschner et al., 2018). NTS are traditionally composed of cognitive and intrapersonal skills that are required for surgical performance, and they are often categorized into constructs such as decision-making, situation awareness, communication, teamwork, and leadership (Flin et al., 2015; S. Yule, Flin, Paterson-Brown, et al., 2006). Several studies have reported positive correlations between increase in communication breakdowns, and the number of surgical errors or reported incidences (Gawande et al., 2003; Lingard et al., 2004; Panesar et al., 2012). With the increased awareness that these skills of surgical team members, especially those of the surgeons, contribute critically to surgical outcomes, several assessment tools have been developed.

Current NTS assessment tools are largely check-list based evaluations. The behavior rating systems such as the Non-technical Skills for Surgeons (NOTSS), Oxford Non-Technical Skills (NOTECHS), and Observational Teamwork Assessment for Surgery (OTAS) are used by expert raters to evaluate surgical team members (e.g., attending surgeon, anesthesiologist, nurse) or the entire surgical team (Mishra et al., 2009; Undre et al., 2007; S. Yule et al., 2008). These tools are comprised of elements that describe a construct, which are then used to calculate an overall NTS score. Although the elements and behavioral rating scales vary for each tool, the assessments are generally similar in that scores are assigned to each element, and a higher score represents an individual or team having higher NTS. For example, communication and teamwork (as one construct) in the NOTSS evaluation tool is described by the elements of exchanging information, establishing a shared understanding, and coordinating team activities. A rater evaluates these elements of a surgical team member throughout the procedure and assigns a single representative score between 1 (poor communication/teamwork) and 4 (good communication/teamwork). These communication and teamwork element scores are used to calculate a construct score and all construct scores are used to compute an overall NTS score for the surgery.

These evaluation tools have been the standard method of NTS assessment; however, many limitations exist with the current measurement system. Primarily, the behavioral rating requires a trained person to rate NTS. While self-assessments have been studied, these tools are primarily used by trained experts, which are time-intensive in terms of both training and the evaluations

themselves. Furthermore, raters' assessments are subject to their own biases, which increase the intra- and inter-rater variability, making obtaining consistent results difficult. Therefore, there is a need for objective measurements of NTS. As NTS are skills that are needed for the efficiency of performance, task performance metrics may be a form of objective metrics of NTS. However, these metrics are also limited in that they are obtained post-hoc. To address the gap of NTS measurement and assessment, implementation of objective metrics that can be measured in real-time are needed.

1.1 Significance

NTS assessments in the operating room (OR) are currently limited to observational, rater-based assessments. To overcome the limitations of these subjective tools, an objective measurement is needed. This work aims to take the first steps in identifying objective NTS metrics for clinicians and applying such metrics for NTS evaluation in the clinical environment. The identification of quantitative behavioral or physiological metrics of NTS can provide a basis for future studies for more objective evaluations.

With the increased knowledge of objective measures and sensor technology, NTS assessment can be expanded for synchronized evaluation of the entire surgical team and to provide immediate feedback of individuals or teams being evaluated. Obtaining NTS metrics can be automated with the use of unobtrusive sensors that can be worn by individuals in addition to environmental sensors placed in the OR to detect all surgical team members. The joint implementation of sensors and detection algorithms for NTS behaviors can lead to near real-time assessment to help immediately identify NTS and possibly indicate to the clinical team of potential NTS issues. The advancement in evaluation techniques can potentially lead for better training and evaluation of NTS for not only trainees but for experts as well.

Furthermore, more complex models of NTS can be found with objective metrics. Due to the reliance of raters, current NTS constructs are limited in behaviors that can be observed or inferred. With the use of objective physiological metrics, deeper understanding of additional skills, but particularly cognitive skills, can be obtained. The inclusion of additional metrics can provide further insight for a holistic understanding of these skills and can be used for the development of more comprehensive models of NTS.

1.2 Research Problem

Objectively measuring NTS in surgery includes (a) understanding NTS objective assessments, (b) implementing such metrics, and (c) validating the objective-based assessment with current assessment techniques. This dissertation tries to address the following research questions to understand objective measures of NTS in surgery.

1.2.1 Research Question 1 (RQ1)

What are metrics that measure NTS objectively?

There is a limited understanding of the possibility of using objective metrics to measure NTS in clinical environments. Literature on the topic has been dispersed across many disciplines such as medicine and engineering. Answering this question will provide a map of the literature on objective metrics to assess clinicians' NTS. The consolidation of current objective measurement techniques in the literature can help build a foundation in further understanding NTS beyond assessments based on behavioral markers.

1.2.2 Research Question 2 (RQ2)

Can clinicians' NTS be measured objectively in clinical environments?

Having identified objective metrics in RQ1, the implementation and validation of the metrics in measuring NTS in surgery needs to be completed. The objective metrics will be used to predict current reference-standard NTS measurements (i.e., rating-based tools) in the clinical environment, specifically in an OR, to understand NTS of surgeons. This can further the understanding of NTS characteristics and of directional relationships of the objective metrics with the specific behaviors.

1.3 Summary of Document Structure

This chapter provides an introduction to the motivation for objective NTS measurement and states the two research questions that this dissertation aims to answer. Chapter 2 is the literature review elaborating on the current knowledge and methodologies of NTS constructs, assessments, and select physiological and behavioral metrics. Chapter 3 presents the scoping review of the literature to answer RQ1, while Chapter 4 describes implementation of objective metrics in the OR, answering RQ2. Chapter 5 provides a general discussion of this work, with proposed

guidelines to address limitations from this work. Finally, Chapter 6 summarizes conclusions from this dissertation.

2. LITERATURE REVIEW

Individual and team skills that affect patient outcomes in clinical environments have been identified and studied since the 1980s (J. B. Cooper, 1984; Gaba, 1989). These studies stemmed from the identification of human error in anesthesia and documented human behaviors and factors that affected decision-making and anesthetic vigilance and monitoring in the operating room (OR) (Biebuyck et al., 1990; Gaba, 1989). Helmreich and Schafer (1994) identified parallels of team performance in the OR to those of aviation, and they called for formal training in human-factors aspects that affects team performance (e.g., crew resource management; CRM) to reduce human error. Literature on the constructs of non-technical skills (NTS; e.g., communication, teamwork, decision-making, leadership) and human factors in medicine were expanded since 2000, when the Institute of Medicine published the "To Err is Human" report. This landmark report called for efforts to improve patient safety in health care, which included addressing aspects of NTS (Baker et al., 2005; Baldwin et al., 1999; Hu et al., 2016; Kiekel et al., 2017; Kohn et al., 2000; Lingard et al., 2004; Lingard & Haber, 1999; Rosen et al., 2008; Weaver et al., 2010).

Interpersonal and cognitive skills that influence technical skills, as well as the safety and efficiency of operations, are the focus of NTS assessments of surgeons (Flin & O'Connor, 2017). These skills – or constructs – revolve around behaviors of individuals used and observed in the surgical workplace. Although some frameworks and taxonomies of intraoperative NTS include additional cognitive skills (e.g., cognitive workload management, coping with stress and fatigue), this work will focus on metrics that have been evaluated through current assessment tools through observed behaviors (Flin & O'Connor, 2017; S. Yule et al., 2008). The following sections will describe NTS constructs, common assessment tools, and selected objective measures that are relevant in objective measures of NTS in surgery.

2.1 Constructs

The five common constructs of NTS used in the assessment tools are described below: communication, teamwork, leadership, situation awareness, and decision-making. Due to the broad nature and understanding of each construct, the following section will review selected theory

behind the construct, its application in the literature, and examples of assessment tools that refer to or measure the specific construct.

2.1.1 Communication

Communication is critical for successful team performance and shared cognition (E. E. Salas & Fiore, 2004). It is defined as the transmission of information between two or more individuals in the team (Flin & O'Connor, 2017). Elements of communication include exchanging or requesting information or instructions so that the sender and receiver can encode and decode the meaning of the information, respectively. Communication is often modeled by one-way or two-way communication; however, although both communication types occurs in surgery, it is often emphasized that two-way communication is more accurate. This is due to the presence of feedback by the receiver in the latter communication model, as shown in Figure 2.1.

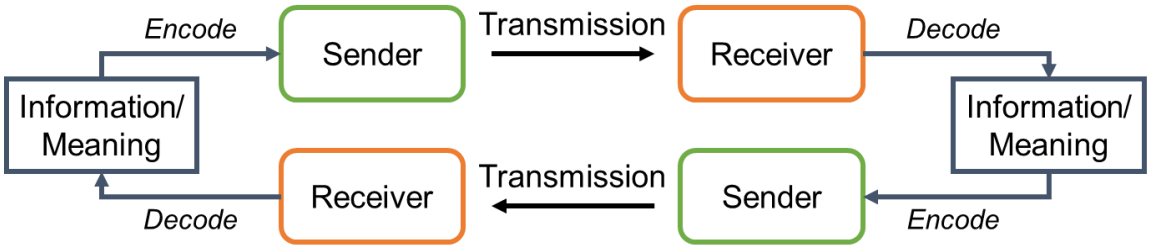

Figure 2.1 Model of two-way communication adapted from (Flin & O'Connor, 2017)

In general, theories on communication are vast in that there is a specific discipline exploring this construct. In verbal communication, language is used to complete a joint task by coordinating members in a team (Clark, 1996). In the act of communication, a person is either a speaker or a listener. To perform a team task, the speaker(s) and listener(s) cooperate by exchanging an utterance to convey their intensions (G. A. Miller & Glucksberg, 1988; Proctor & Van Zandt, 2008). An utterance can be described by the semantics – the meaning of the words – or the pragmatics – language used in a context. In addition to the meaning or language of spoken words, different strategies for establishing common ground – or mutual knowledge or understanding – between speakers and listeners have been used to describe information and logic of a conversation. These includes frameworks such as the Given-New Strategy and Cooperative Principle (Grice et

al., 1975; Haviland & Clark, 1974). The Given-New Strategy describes the integration of information that the listener is expected to previously know (Given) with new information (New) that is given by the speaker (Haviland & Clark, 1974), while the Cooperative Principle focuses on maxims of quantity, quality, relevance, and manner for effective information sharing to establish the common ground (Grice et al., 1975). Understanding these theories and applying their principles in team settings, such as surgery, can lead to more efficient information transfer and communication. In an OR, team communication has been described to follow the rhetorical theory of communication, where the speaker convinces the listener with a message (Lingard et al., 2002). Additionally, it has been further described that communication is driven to achieve a common ground for efficient exchange, and this requires recognition and negotiation of shared interests (Lingard et al., 2002).

Non-verbal communication describes communication by means other than words (Knapp et al., 2013) and includes variables of face-to-face interaction such as prosodic features of speech and other body expressions or gestures. These behaviors include eye movements (e.g., gaze behaviors), pointing, head nods, posture, and proximity between the speaker and listener (Duncan & Fiske, 2015). Non-verbal communication is often utilized between team members to complete teamwork in the environment.

In evaluating individual or surgical teams, prosodic elements are often grouped as non-verbal communication (Malandro & Barker, 1983), as they are reliant of language. Prosody is used to describe auditory properties of speech production, transmission, and perception during the transfer of messages (J. Fletcher, 2010; Wennerstrom, 2001). Prosodic features can be described in acoustic and auditory measures, where acoustic features describe the physical (e.g., time and spectral components) properties of a sound wave, and auditory features are the subjective impression of the acoustic feature to a listener. Common acoustic metrics include fundamental frequency (Hz), duration (time), intensity (dB), and spectral characteristics such as energy (dB). In addition, auditory metrics are often described by pitch (Hz) and loudness (dB). Other features used to describe speech include tempo, which includes articulation rate (syllables/s) and the number of syllables uttered per second minus pauses (J. Fletcher, 2010).

Non-verbal modes of communication have been investigated in several clinical settings. These behaviors have been observed and examined in interdisciplinary clinical teams, where studies note that non-verbal cues go in-hand with good leadership behavior and situation awareness

of team members (Härgestam et al., 2016; Mitchell et al., 2011; Moore et al., 2010). Furthermore, studies have integrated these communication techniques to measure the communication construct between surgical team members during surgery, where the frequency of the non-verbal cues for requesting or passing instruments were observed and quantified (Korkiakangas et al., 2014; Tiferes, Bisantz, et al., 2016).

Strategies for communication have been developed and trained in the clinical environment, such as closed-loop communication and Situation-Background-Assessment-Recommendation and Request (SBAR), in order to reduce ambiguity and facilitate efficient exchange of information between team members (Härgestam et al., 2013; Kesten, 2011). In addition, elements of these strategies have been assessed using specific communication evaluation tools such as the Communication and Teamwork Skills and Pediatric Resuscitation Communication (Frankel et al., 2007; Parker-Raley et al., 2012; Rehim et al., 2017). These behavioral assessment tools include communication elements within the team turn-taking (e.g., letting others speak without interruption, not talking over people) and the emphasis on information transfer after training programs to improve NTS (Gillespie et al., 2010).

2.1.2 Teamwork

Teamwork is described as two or more individuals working together in a team to perform a task or complete a goal. Elements of teamwork in NTS frameworks include establishing a shared understanding; coordinating team activities; and understanding team needs (Paris et al., 2000; S. Yule, Flin, Paterson-Brown, et al., 2006). Team working and performance is also described by the process or actions of coordination, cooperation, or back-up behavior, which are also used as constructs in different NTS assessment tools. Teamwork in a surgical setting has been focused on measuring, managing, and training team performance.

In their review of teamwork, Paris et al. (2000) summarized theoretical contributions to the concept of teamwork and described three categories of teamwork dimensions based on specific competencies (e.g., mutual performance monitoring, collection orientation; exhibiting flexibility): cognition, behavior, and attitudes. These theories were representative of different approaches such as the social psychological, sociotechnical, ecological, and human resource and technological approach. Team performance has also been modeled as a process model: individuals and environments as inputs, dynamics of the inputs as throughput, and performance as output

(Unsworth & West, 2000). Individuals hold different positions (e.g., leader or member) within a team and come together with their own knowledge, skills, and attitudes to an organization to complete the task (e.g., surgery). Processes or dynamics are described as communication, coordination, and cooperation that occur for team cognition, or how the team "thinks". This includes not just an individual team member's mental processes but the interactions between the members (Cooke et al., 2004; Fernandez et al., 2017). Outputs of teams are signified by their performance, measured through metrics such as productivity or the number of errors or accidents. This framework reflects NTS elements for actions that are needed to increase team cognition and performance within surgeon assessments.

Team behavior processes also reflect concepts of individuals' and teams' adaptability, mutual performance monitoring, collective orientation, and shared vision (Flin & O'Connor, 2017). Measures of these behaviors and processes have been described by dimensions such as providing feedback, closed-loop communication (i.e., verifying and confirming intended communication was received), and backing-up behaviors for CRM (E. Salas et al., 2000). These elements have also been adapted in NTS frameworks in the OR for teamwork assessment of an individual or entire team, focusing on behaviors for effective joint team task completion and performance.

Several studies have investigated teamwork in healthcare teams, which have included assessment through self-, expert-, and peer-assessment (Baker et al., 2010; Makary et al., 2006). For example, the TeamSTEPPS® Teamwork Attitudes Questionnaire has been adapted to evaluate an individual's self-perceived attitude of teamwork (Sawyer et al., 2013; Watanabe et al., 2019). This questionnaire is comprised of a Likert scale that asks a team member to assess their impressions on team structure, leadership, situation monitoring, mutual support, and communication (Baker et al., 2010). Results from this tool is often used to evaluate the changes of an individual's attitude toward teamwork after team training (e.g., TeamSTEPPS®), while team performance ratings through rater observation can be also completed (Sawyer et al., 2013; Zhang et al., 2015). Additionally, Makary et al. (2006) evaluated teamwork by asking surgical team members to complete the Safety Attitudes Questionnaire adapted to the OR. Ratings from the individuals were used to assess perceived amount of good collaboration among different team member roles. It was found that there was a disconnect between surgeons and nurses: surgeons reported 88% perceived high quality of collaboration and communication with OR nurses, while nurses reported this only 48% of the time (Makary et al., 2006). With the ability of these

questionnaires to be administered by individuals to evaluate not only self, but team member's as well, teamwork can be leveraged to gain insights on the teams' strengths and ability to effectively investigate and improve team interaction and skills.

2.1.3 Leadership

In team performance, leadership is defined as the process of influencing the activities of individuals or a team to accomplish a goal (Hersey et al., 1979; Hjortdahl et al., 2009). Elements describing leadership include setting and maintaining standards, supporting others, and coping with pressure. Leaderships goes in-hand with aspects of teamwork, as a leader is an individual with a specified role that is integrated with the team and brings his/her own knowledge, skills, attitudes, and leadership style. NTS assessment of leadership favors a horizontal leadership approach, where the surgeon is not recognized as the automatic leader of the team, and there is emphasis on all team members participating in the decision-making process (Gostlow et al., 2017).

There are several theories of leadership but transformational, transactional, and passive leadership is used primarily to describe team leaders (Avolio & Bass, 2004; Bass, 1997; Horwitz et al., 2008). Transformational leadership is described with characteristics of a leader working with teams to enhance motivation and behaviors of team members. Transformational leaders connect with the followers' needs and encourage developing them so that the followers' performances exceed expectations. On the other hand, transactional leadership focuses on teamwork through a reward and punishment system. Transactional leaders are task oriented and focused on management by exception (e.g., mistakes and failures) (Hu et al., 2016). Finally, passive leadership is described as passive management by exception (e.g., avoiding action until mistakes and failures can no longer be ignored) and laissez-faire, or delegative leadership where leaders are hands-off and allow followers to make decisions.

Questionnaires have been developed to assess leadership in the surgical environment. The Operating Room Management Attitudes Questionnaire elicits attitudes regarding leadership style and describes them in four ways: autocratic, mild, consultative, and democratic. Autocratic describes the most vertical relationship (i.e. leader is superior and expects obedience and no question) while democratic is represented by horizontal distribution of power (i.e., puts problem before group and invites discussion) (Helmreich & Schaefer, 1994; Schaefer & Helmreich, 1993). The Multifactor Leadership Questionnaire is commonly used to self-assess or observe

transformational and transactional leadership through 45-itemed survey. The adapted version for surgeons includes assessment of observed intraoperative behaviors that are expected to impact patient safety and team performance. The elements describing leadership in the assessment tool include those that are described in general NTS assessments but include elements such as directing and training others. Moreover, the Surgeons' Leadership Inventory was more recently adapted from the Multifactor Leadership Questionnaire to describe a taxonomy and behavioral measurement tool for intraoperative leadership skills (Avolio & Bass, 2004; Parker et al., 2012, 2013). Previous work has used these tools to quantify leadership styles of surgeons in the clinical settings. From video recordings of surgeons, Hu et al. (2016) concluded that increased behaviors of transformational leadership improved team performance, as defined by increases in information sharing and supportive behaviors; however, in a different study of surgeon observations, Barling et al. (2018) did not find this positive relationship but reported a significant inverse relationship between negative leadership behaviors and team performance. These results show the continued need for identification of leadership behaviors strategies to enhance the effectiveness of surgical teams (Sadideen et al., 2016; Stone et al., 2017).

2.1.4 Situation Awareness

Situation awareness and vigilance as an NTS are centered around perception and attention. The construct is composed of the elements of gathering information, understanding information, and projecting and anticipating future state. The need to measure situation awareness in healthcare started in the 1980's: monitoring, awareness, and vigilance were identified as behaviors needed by anesthesiologists to reduce human error that were associated with substantive negative outcome (e.g., intravenous drug overdose, wrong drug, and inappropriate ventilation) (J. B. Cooper et al., 1984). As a cognitive skill, descriptions and the respective evaluation of the construct elements are centered around observable behaviors. An example of a good behavior for anticipating future state includes a surgical team member planning an operating list to account for potential delays or challenges and verbalization of what may be required later in the operation. An example for anesthesiologists displaying poor situation awareness is reducing the level of monitoring because of distractions; moreover, an example of a surgeon with poor awareness is informing the team only after encountering predicable blood loss (S. Yule, Flin, Paterson-Brown, et al., 2006).

Situation awareness considers memory and comprehension in complex, dynamic systems (Proctor & Van Zandt, 2008). Although there are several definitions and theories of situation awareness, it is commonly defined as "the perception of the element in the environment within a volume of time and space, and the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988a, p. 97). This definition follows Endsley's model describing three levels: perception of the elements, comprehension of situation, and projection of future states. Furthermore, Bedny and Meister (1999) defined situation awareness through theory of activity, where three stages (i.e., orientation stage, executive stage, and evaluative stage) are involved in the current situation motivating an individual to perform an action toward achieving a goal (Salmon et al., 2009). NTS assessment tools in healthcare tend to follow Endsley's model of generating a mental model – a representation of an individual's interpretation of a system or of the world – of the situation through perceiving the status (level 1) and comprehending the environment (level 2). The generation of possible future states (level 3) assumes that individuals develop and maintain mental models of how something works in the environment through training and experience to complete the relevant task. This construct has also been expanded to both shared and team situation awareness, where shared situation awareness describes the overlap of elements between team members while team situation awareness is the degree of situation awareness each team member need to complete their duties (Salmon et al., 2009).

Different techniques have been developed to measure situation awareness. These include situation awareness requirement analysis, freeze/real-time probes, self-rating or observer-rating, performance measures, and physiological measures (e.g., eye tracking) (Salmon et al., 2006). Salmon and colleagues (2006) reviewed the different assessment and calls for the development of a novel approach or a combination of techniques to obtain the most successful situation awareness measurement. It is noted, however, that the Situation Awareness Global Assessment Technique (SAGAT) developed by Endsley (1988b) is arguably the most common approach for this measurement. Furthermore, there has been increased attention on leveraging physiological signals to measure situation awareness. Cardiovascular measures (e.g., heart rate variability) has been correlated with situation awareness in non-healthcare domains such as automotive and aviation (Kunze et al., 2019; Sun et al., 2017). Additionally, the use of eye-tracking technology have been proposed to measure gaze behavior to infer situation awareness, however, the "look-but-failed-to-see" phenomenon is noted as a limitation (Brown, 2002; Salmon et al., 2006).

2.1.5 Decision-Making

Decision-making is defined as the process that a person "reason about and choose between different actions" (Proctor & Van Zandt, 2008). In NTS assessments, elements include considering options, selection and communication of the options, and implementing and reviewing the decision (S. Yule et al., 2008). This cognitive process is also inferred from observable behaviors of actions of verbalizations such as discussing options with team members, clearly communicating a decision, and reconsidering an option if a problem occurs (Flin et al., 2003, 2015).

There are two ways to describe a person's decision making: normative and descriptive. The normative decision-making model describes what people should consider in an ideal circumstance to make the best decision possible. However, descriptive decision-making theory looks at how people really make decisions and include how people overcome cognitive limitations and biases (Lehto & Nah, 2006; Proctor & Van Zandt, 2008). For individuals within dynamic environments, such as surgery, four principles of decision-making have been identified: intuitive, rule-based, analytical, and creative (Flin et al., 2007; Orasanu & Fischer, 1997). Intuitive decision-making is also referred to as recognition-primed decision making (Klein, 1993) or "thinking fast" (Kahneman, 2011). This describes using stored patterns or memories to infer the thinking "When X happens, I know to do Y". Rule-based decision making is consciously searching a person's memory to retrieve the learned match rule for the situation in a procedural manner, "If X, then Y". Analytical decision making is referred to as "thinking slow" (Kahneman, 2011) and refers to the consideration of different options to select an optimal solution ("If X, then Y. But if X and Q, then Z is better"). Finally, creative decisions will devise new course of actions and are often used to propose solutions to unfamiliar problems ("If X, then it could be A, B, or C.") (Flin et al., 2015). These theories have been applied to model surgical decision-making in the OR. Cristancho et al. (2013) developed a naturalistic model of intraoperative decision-making that included elements of assessing the situation, reconciliation cycle. and implementing the course of action; furthermore, Madani et al. (2017) developed a conceptual framework of cognitive processes for intraoperative decision-making to further understand advanced cognitive functions (e.g., planning or error recognition) needed for expert surgical performance.

There are limited decision-making assessment tools in a healthcare setting. Current assessment tools include questionnaires evaluating patient-doctor shared decision-making and informed decision-making (Braddock et al., 2008; Kriston et al., 2010); however, intraoperative

decision-making is typically completed within an overall NTS assessment alongside evaluation of other cognitive or interpersonal constructs (Robertson et al., 2014; S. Yule et al., 2008). Moreover, surgical decision-making has been evaluated in simulation settings (Barzallo Salazar et al., 2014; Leff et al., 2017). These studies identified decision points that are critical in a scenario or procedure and measured an individual's response through the presence of an action or inaction (Barzallo Salazar et al., 2014) or brain activation (Leff et al., 2017).

2.2 Current NTS Assessment Tools

Assessment tools were developed to evaluate NTS of surgical team members to improve team performance. A recent systematic review by McMullan et al. (2020) identified 31 observational tools that quantify NTS. Such tools include the Observational Teamwork Assessment for Surgery (OTAS), Non-technical Skills for Surgeons (NOTSS), Oxford Non-Technical Skills (Oxford NOTECHS), and the Anaesthetists' Non-technical Skills (ANTS), which were all independently created as observational measures to quantify NTS (G. Fletcher et al., 2003; Healey, 2004; Mishra et al., 2009; S. Yule et al., 2008). Authors of these assessment tools identified relevant behavioral markers through task analysis, expert interviews, and observations, and each identified their own constructs and elements to assess. Although there are overlaps of constructs among the tools, the description of the elements describing a construct (e.g., gathering and understanding information, and projecting and anticipating future states for situation awareness) are varied. Figure 2.2 shows the constructs that are included in the assessment tools described below. Yule et al. (2006) noted that these behavior rating systems should be comprised of observational behaviors that contribute to performance; encompass the most important behaviors; use domain-specific language and terminology; and be explicit and reliable. It was noted that social skills should be directly observable and that cognitive skills should be inferred from observing behaviors. The Non-TECHnical Skills (NOTECHS) system to assess CRM skills that are translated from aviation to surgery builds on the described observable criteria: communication is not an individual construct because "communication skills are inherent in all four [constructs] and the listed behaviors all involve communication" (Flin et al., 2003, p. 99).

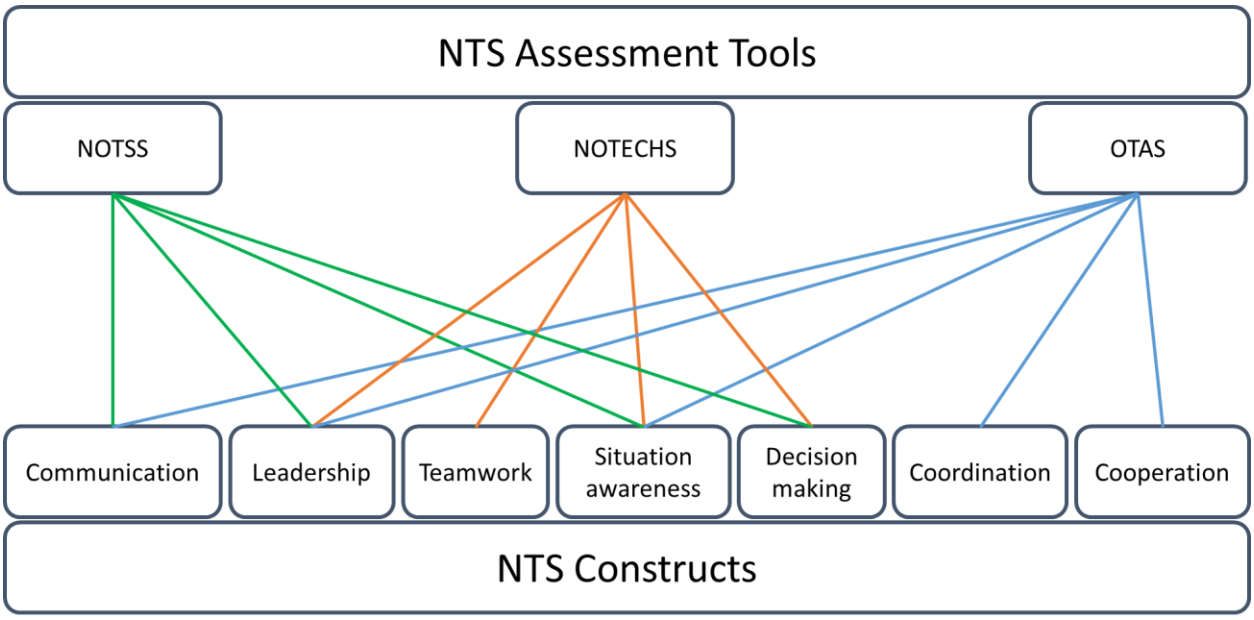

Figure 2.2 Relationship of selected NTS assessment tools and NTS constructs

Moreover, NTS assessment tools have been adapted for different NTS constructs, roles, and specialties (G. Fletcher et al., 2003; Frankel et al., 2007; Mitchell et al., 2012; Steinemann et al., 2012). For example, construct specific assessment tools include those that evaluate leadership and teamwork: the Surgeons' Leadership Inventory was developed to evaluate surgeons' leadership behaviors (Parker et al., 2013) while the TeamSTEPPS® Teamwork Attitudes Questionnaire has been used for self-assessment of teamwork among the clinical team (Baker et al., 2010). Furthermore, evaluation tools for surgical roles have been used, such as the ANTS for anesthesiologists and the Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLITS) for scrub nurses (G. Fletcher et al., 2003; Mitchell et al., 2012). The previously stated assessment tools have been used to evaluate trainees such as medical or nursing students; in addition, tools have been specifically designed for these populations within simulation environments (Cha et al., 2019; Gordon et al., 2019; Moorthy et al., 2006). Furthermore, there has been an increase in customized assessments for various specialties. In particular, literature exists describing tools developed for medical emergency care, ophthalmic care, and surgical techniques such as robotic surgery (S. Cooper et al., 2010; Raison et al., 2017; Steinemann et al., 2012; Wood et al., 2020).

2.3 Objective Metrics

Although skill assessments of surgical teams are still largely rater-based, other techniques have been proposed for intraoperative skills assessment (Dias et al., 2018; Moorthy et al., 2003). Studies have measured motion, trajectory, and force of surgical instruments manipulated by surgeons alongside their gestures for objective surgical skills evaluation (Ahmidi et al., 2013; Reiley et al., 2011). Additionally, objective metrics of individual NTS constructs, but particularly of cognitive workload, in several domains have utilized physiological and behavioral measures (Charles & Nixon, 2019; Dias et al., 2018; Kazi et al., 2019). This work leverages previously established relationships between intraoperative skills and objective measures to initially identify such metrics to quantify NTS. In the below sections, selected objective metrics will be described: speech, interaction, eye-tracking, brain activity, and heart rate variability. Each section will describe the metric, physiology or behavior relevant to the metrics, and examples of analysis previously reported to measure social or cognitive skills.

2.3.1 Speech Metrics

Speech metrics in this work is a broad term encompassing variables obtained from the verbal communications during a spoken utterance. In addition to the lexical content of speech (i.e., what is said), speech metrics include prosody patterns, or how something is said (Rosenberg, 2009). Prosodic patterns of speech include pitch, duration, stresses on specific words, and pauses. Signal processing of speech and a summary of the fundamentals of these processes can be found in literature (Santen et al., 2008).

Speech production is completed through the two organ groups: phonation and articulation. The phonatory organs are composed of the lungs and larynx to create voice sounds through changes in air pressure and vocal vibration, which control prosodic speech patterns such as pitch and loudness. The articulation organs are comprised of the lower jaw, tongue, lips, and velum that control resonance and modulations of the voice to generate certain sounds. The two organ groups work together to produce vowels and consonants (Santen et al., 2008).

There has been work leveraging both lexical and prosodic patterns to assess communication data. One such technique includes latent sematic analysis, which has been used to automatically measure team communication and cognition (Gorman et al., 2003; Kiekel et al., 2001). This

approach uses transcriptions of speech to measure the semantic similarity between strings of text (Landauer et al., 1998). It has also been used to obtain measures such as communication density (i.e., the relationship between the meaningfulness of a discourse and the number of words spoken) and lag coherence (i.e., turn-taking of utterances). Additionally, general turn-taking variables such as turn time, turn mean length, interruption rate, and pause rate are used to describe communication (Duncan & Fiske, 2015).

Analysis of prosodic patterns has been the approach taken for semi-automatic audio analysis (Peng et al., 2019). Studies have found the association between fundamental frequency, tone, and intensity with politeness (Grawunder & Winter, 2010; Laplante & Ambady, 2003; Loveday, 1981). Other studies have found relationships between prosodic patterns and stress, which found different prosodic patterns (e.g., pause duration and articulation rate) for leaders and experts compared to followers and learners (Protopapas & Lieberman, 1997; Scherer et al., 2012).

2.3.2 Interaction Metrics

Interaction metrics describe non-verbal actions and communication that occur during face-to-face interaction. These include gestures and gaze in addition to a more global consideration of how far two speakers are away from each other and whether they are facing each other. Face-to-face variables considered in Duncan & Fiske (2015) include cues such as nod rate, short/long vocal back-channel rate (e.g., short responses such as “I agree”), smiles, gaze behaviors, gesturing, self-adaptors (e.g., touching accessories or parts of body), foot movements, and postural shifts. The anatomy of these interaction metrics ranges from the movements of muscles such as the zygomaticus major and minor to pull up the corners of the mouth for smiles to overall body positioning and location of a person relative to other objects in a room.

Algorithms and technologies exist to measure these metrics. Advances in computer vision has allowed smiles to be automatically detected from images and video (Bousmalis et al., 2009; Whitehill et al., 2009) and eye-tracking technology has allowed for the analysis of gaze behavior, which will be discussed further in a following section. Additionally, body movements can be captured through unobtrusive motion capturing systems to measure the number of a specific gesture. The position of people and objects has been quantified through positioning sensors to determine the location of a nurse in a simulated OR to build patterns for team member visualization during a procedure (Echeverria et al., 2018). For the entire surgical team, Ahmad et al. (2016) has

investigated surgical team members' movements and dynamics during robotic-assisted surgeries and classified their movements and locations for potential improvements of an OR layout and workflow.

2.3.3 Eye-Tracking Metrics

Eye-tracking metrics include visual scan patterns of the environment and how the eye physiology changes with stimuli. Metrics focused on conscious eye movements reflecting the visual attention in the environment include number and duration of fixations, gaze points, areas of interest, and number of saccades (i.e., rapid eye movements between fixation points). Involuntary pupillary response includes metrics such as pupil dilation and blink rate. Current eye-tracking equipment uses the pupil center corneal reflection to detect and measure the changes in eye movement and project gaze, or visual field, maps. Saccades and fixations, or stops, are quantified to determine patterns of interest.

Conscious and unconscious eye responses have been used to differentiate expertise and surgical skill in many domains (Gegenfurtner et al., 2011; T. Tien et al., 2014; Wu et al., 2019). A systematic literature review by Tien et al. (2014) concluded that it is feasible to use eye-tracking metrics and gaze-tracking to differentiate surgical skills based on expertise (e.g. experts and trainees). These findings were also reflected in not just the surgical domain: a meta-analysis of eye-tracking metric differences among expertise levels in professional domains (e.g., sports, medicine, aviation, and driving) found shorter fixation duration, more fixation on task-relevant areas, and longer saccades in experts compared to non-experts (Gegenfurtner et al., 2011). Eye-metrics have also been related to surgical skill defined not only by expertise level but also performance metrics such as completion time and accuracy in surgical training (Sodergren et al., 2011; Vine et al., 2012). Assessments of gaze-training (i.e., training mechanism to allow participants to expert gaze patterns) and orientation-training (i.e., tutorials identifying regions of interest and exemplary gaze fixation sequences) in surgical simulation reported better performance among participants who received the training than those who did not. In addition to assessments of technical skills, eye-tracking metrics have been related to cognitive skills.

Eye metrics have been associated with cognitive activity, workload, situation awareness, and decision-making (Al-Moteri et al., 2017; Gegenfurtner et al., 2011; Haapalainen et al., 2010; Marshall, 2002; Pomplun & Sunkara, 2003; T. Tien et al., 2014; Wu et al., 2019). In laboratory

environments, pupil characteristics such as dilation and number of blinks have been found to be associated with emotion, arousal, and stress in various domains (Hess, 1965; Marshall, 2000; Partala & Surakka, 2003; Pomplun & Sunkara, 2003).

In addition to measuring these cognitive and physiological states, eye-tracking metrics have been used in the clinical environment to measure cognitive load, situation awareness, and decision-making (Desvergez et al., 2019; Dias et al., 2018; O'Meara et al., 2015). For example, eye-tracking measures were associated with self-perceived workload, especially with gaze entropy (i.e., the randomness of visual scanning) within robotic-assisted surgery skills training (Wu et al., 2019). Additionally, measures of situation awareness through questionnaires and eye metrics have been used in simulated clinical settings. O'Meara et al. (2015) used eye-tracking technology to obtain gaze fixation of nurse trainees in a clinical scenario to infer their situation awareness through the eye-mind hypothesis, that the eye fixates on what the mind is focusing upon (Duchowski, 2007; Just & Carpenter, 1975). The authors then used the gaze patterns to aid in providing feedback to nurse trainees to improve their situation awareness. This measurement of situation awareness through questionnaires and eye-tracking measures have also been completed for other clinical team roles such as anesthesiologists (Desvergez et al., 2019; Grundgeiger et al., 2015), where studies map the gaze patterns and measure the monitoring of critical equipment. Finally, Al-Moteri et al. (2017) completed a scoping review of literature on the usage of eye-tracking for medical decision-making, focusing on visual cue processing. Relevant articles were synthesized to identify three errors related to decision making: detection, recognition, and judgmental error. Although eye-tracking was identified to be associated with decision-making, there was a recognition that further work should be completed to evaluate analytical decision-making with eye tracking.

2.3.4 Brain Activity Metrics

Technologies that measure brain activity and function in the field of neuroergonomics have been utilized in application areas such as mental workload, vigilance measurement, and brain-computer interfaces (Parasuraman & Wilson, 2008). These neuroimaging technologies often measure brain blood flow (e.g., functional magnetic resonance imaging [fMRI], functional near-infrared spectroscopy [fNIRS], or transcranial Doppler [TCD] sonography) or measure neural activity (e.g., electroencephalography [EEG]). Metrics that are commonly obtained and reported from these signals include measurements from the acquisition systems (e.g., hemoglobin

concentration from fNIRS or spectral power from EEG) or measures of activation from brain regions. In medical simulation and clinical environments, fNIRS and EEG systems have been used to measure cognitive skills.

Objective surgical skills assessment has leveraged fNIRS to model brain behavior in simulation (Leff et al., 2008; Nemani et al., 2019, 2017). Studies have used this technology to discriminate expertise (e.g., trainee and experts). Leff et al. (2017) found activation in the dorsolateral prefrontal cortex in experts during surgical decision-making which was not observed in trainees; furthermore, Nemani et al. (2017) found increased functional activation in experts compared to novices in prefrontal cortex channels. The latter study also concluded that differentiating expertise levels with technical performance and brain activity combined is more sensitive than using just performance score. Additionally, changes in prefrontal activation have been investigated in relation to motor skills. While assessing surgical skill transfer, it was found that fNIRS is more accurate in classifying surgical motor skill transfer than performance metrics measured through subjective performance checklist assessment and completion time (Nemani et al., 2019). In addition, another study also found a relationship between technical performance and prefrontal activation, and activations in this and the anterior cingulate cortex were related to changes in task familiarity for surgical novices and differed from expert surgeons while performing a simulated surgical task (Leff et al., 2008).

To measure the desired brain activity of different regions, EEG equipment measures electrical activity obtained from electrodes placed on the scalp. Common EEG metrics obtained from recordings are rhythmic activity divided into frequency bands: delta (< 4 Hz), theta (4 – 7 Hz), alpha (8 – 15 Hz), beta (16 – 31 Hz), and gamma (> 32 Hz). (E. K. Miller & Cohen, 2001; Sauseng et al., 2005). Depending on the technology, there can be between 10 to 82 channels placed on the scalp that are used to detect different brain patterns. Previous literature from many domains have utilized EEG to analyze attention, cognitive workload, fatigue, and engagement; moreover, additional analysis relating these constructs have been associated with different regions of the brain such as the frontal and parietal activation with tasks manipulating cognitive skills and working memory (Berka et al., 2004, 2007; Borghini et al., 2012; Klimesch, 1999; Klimesch et al., 1998) In addition, different EEG systems have been used in the in the OR, specifically exploring cognitive workload changes of surgeons (Carswell et al., 2005; Guru, Esfahani, et al., 2015; Guru, Shafiei, et al., 2015). The aspects of perception and cognition measured by EEGs in previous

studies are embedded in the underlying theory of NTS constructs. For example, Miller and Cohen (2001) found prefrontal cortex beta activity were related with retrieval of cues, rules, and goals, which are needed for during a decision-making process while Jacobs et al. (2006) found that theta band power correlated with memory retrieval and decision making.

2.3.5 Cardiovascular Metrics

Cardiac electrophysiology investigates the electrical activities of the heart. Electrical activity when a heart contracts are described with different intervals of the cardiac cycle. In particular, components of an electrocardiogram, which models the voltage from the cardiac muscles over time, are used to represent the cardiac rhythm. The simplified patterns from electrocardiogram are represented as waves and complex (Figure 2.3). The P wave represents the depolarization of the atria, the QRS complex represents the depolarization of the ventricles, and the T wave represents the repolarization of the ventricles (Malik & Camm, 1990; Pomeranz et al., 1985; Pumprla et al., 2002). Deviations from regular patterns can represent disturbances in cardiac rhythm as symptoms of diseases or due to experience from acute events to change heartbeat such as exercise and increased stressed or mental load.

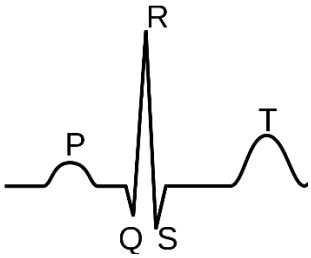

Figure 2.3. Image of the QRS complex. "QRS normal " by A7N8X is licensed under CC BY 2.0

Cardiovascular metrics obtained from electrocardiography such as waveform peaks and patterns of a heart over time have been evaluated in the literature as measures to determine changes in physiological states. Specifically, heart rate variability (HRV) describes the variations of time intervals between heartbeats (Camm et al., 1996). Common HRV metrics include R-R interval (the time between two heartbeats), standard deviation of NN intervals (SDNN; NN representing R-R intervals but emphasizing that the heartbeats are normal), and root mean square of successive differences (RMSSD). In addition to these time-domain variables, the heartbeat data can be transformed to obtain frequency domain metrics. Specifically, the electrical activity generated by

the heart throughout the cardiac cycle are divided into very low frequency (< 0.04 Hz), low frequency (0.04 – 0.15 Hz), and high frequency (0.15 – 0.4 Hz) power. These variables are correlated to the parasympathetic nervous system. For example, the low frequency/high frequency power ratio is often described to represent the balance between the sympathetic and parasympathetic nervous systems (Shaffer & Ginsberg, 2017). These metrics have been applied in the literature as objective metrics, as these variables provide insight on changes in cognitive or physiological states.

HRV metrics have been used to monitor activity that alters the autonomic nervous system, comprised of the sympathetic and parasympathetic stimulation. The sympathetic system responds to stress, and using HRV to measure stress and cognitive workload have been vast in the literature (Castaldo et al., 2015; Dias et al., 2018, 2019; Rajendra Acharya et al., 2006; Veltman & Gaillard, 1996). In a systematic literature review, HRV was found to be the most common real-time measure for quantification of intraoperative surgeon cognitive load; specifically, it was reported that the LF/HF ratio – as it increases with increases in cognitive load – has been frequently used as an objective measure (Dias et al., 2018). Moreover, HRV metrics have been proposed to measure cognitive NTS such as situation awareness. In non-healthcare domains, heart rate and HRV metrics measured through ambulatory monitoring systems (e.g., wearables) have been related to subjective measures of situation awareness (Kunze et al., 2019; Thayer et al., 2009). For example, in a high-stakes scenario-based simulation training environment (i.e., police shooting), significant positive correlations were found between HRV and self-perceived situation awareness as well as with task demands (Saus et al., 2006, 2012). Furthermore, HRV metrics have been associated with executive functions – or processes for controlling behavior – such as reasoning, problem-solving, and planning goal-oriented behaviors in the literature (Hansen et al., 2004; Luque-Casado et al., 2016; Thayer et al., 2009). Thus, this shows the potential of leveraging this objective measure to quantify additional cognitive skills such as decision-making.

2.4 Summary

This chapter summarizes the literature on common NTS constructs, assessment tools, and physiological and behavioral measures that have been used to objectively measure various intraoperative skills. It provides a foundation for the understanding of pertinent concepts to address the posed research questions. This chapter provides a general overview of the relevant concepts,

and the next chapter elaborates on a synthesis of the literature of the current state of objectively measuring NTS of surgeons.

3. OBJECTIVE MEASURES OF SURGEON NON-TECHNICAL SKILLS IN SURGERY: A SCOPING REVIEW

3.1 Introduction

Non-technical skills (NTS) of healthcare teams are critical to patient safety for high-stress, high-stakes environments such as in hospital intensive and emergency care (Gawande et al., 2003; Hull et al., 2012; S. Yule, Flin, Paterson-Brown, et al., 2006), but particularly in the surgical environment where it has been found to affect patient safety (Leuschner et al., 2018). NTS are skills, or constructs, such as communication, teamwork, leadership, situation awareness, and decision-making. These interpersonal and cognitive skills of surgical team members influence technical skills and adverse events in the operating room (OR) (Agha et al., 2015; Hull et al., 2012). For example, evidence of strong relationships between technical error and teamwork failures were stated in a systematic review (Hull et al., 2012), and reports have found that 43% of surgical errors were related to communication (Gawande et al., 2003). In addition, 44% of the incidences reported in orthopaedics and trauma surgery to the National Patient Safety Agency had a failure in NTS (Panesar et al., 2012). With increasing data showing that NTS impact technical skills and patient safety, several assessment tools have been developed to help train and evaluate healthcare workers in NTS.

Centered around a consensus of NTS constructs, many role and specialty-specific assessment tools have been developed to evaluate NTS behavior. A recent review identified 31 observational NTS tools for the OR (McMullan et al., 2020), which included evaluation ranging from the entire team (e.g., Observational Teamwork Assessment for Surgery [OTAS]) down to the individual surgical team members such as the attending surgeon (e.g., Non-technical Skills for Surgeons [NOTSS] and Oxford Non-Technical Skills [NOTECHS]), anesthesiologist (e.g., Anesthetists' Non-Technical Skills [ANTS]), and scrub nurse (Scrub Practitioner's List of Intraoperative Non-Technical Skills [SPLINTS]) (G. Fletcher et al., 2003; Mitchell et al., 2012; Robertson et al., 2014; Undre et al., 2007; S. Yule et al., 2008). These behavior rating systems have been used for NTS evaluation primarily through expert-observation in the OR.

To apply these tools, the team or individual is typically evaluated on point-scales based on various constructs included in the instrument. Often, the constructs are decomposed to different elements with exemplars, and the scores of the construct are used to calculate an overall or average

NTS score. Though the elements and behavioral rating scales vary for each tool, a consensus exists that higher scores are given to behaviors that represent higher NTS that enhances patient safety and effective teamwork (Catchpole et al., 2007). For example, leadership in the NOTSS evaluation tool is described by the elements of 1) setting and maintaining standards, 2) supporting others, and 3) coping with pressure (S. Yule et al., 2008). A rater evaluates these elements of the surgical team member throughout the procedure and assigns a single representative score between 1 (poor leadership) and 4 (good leadership) for each element. These leadership element scores are used to calculate a construct score and all construct scores are averaged to compute an overall NTS score for the surgery. These NTS assessments were specifically developed to measure NTS constructs jointly; however, many training and evaluations exist for measuring specific NTS constructs individually, such as teamwork with the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS^{TM}) and leadership with the Surgeons' Leadership Inventory (Baker et al., 2010; Mayer et al., 2011; Parker et al., 2013).

Although current NTS assessment tools have been the standard method of NTS evaluations, several limitations impact their practice and reliability. Primarily, these observation tools require a trained expert to reliably rate observed NTS, which is time-intensive in terms of both training and the evaluations themselves. Inter-rater reliability has been reported to be in the acceptable range for tools such as the NOTSS (Jung et al., 2018); however, less acceptable reliability has also been reported for constructs such as situation awareness and decision-making between minimally trained raters (S. Yule et al., 2008). This may be due to the increase in variance from inherent biases or errors of raters (Feldman et al., 2012). Additionally, with the difficulty of summarizing hours-long procedures into a single rating score for multiple team members, there is a need for unbiased and time-efficient methodologies to measure NTS.

While NTS assessments in the OR are still primarily observation- and rating-based, many other techniques have been proposed for skills assessment in the intraoperative environment, as well as non-surgical domains (Moorthy et al., 2003; T. Tien et al., 2014). Objective – or quantitative – metrics of technical skills and individual NTS constructs in several domains (e.g., aviation and driving) have primarily leveraged physiological and behavioral measures (Charles & Nixon, 2019; Kazi et al., 2019; Tiferes et al., 2015; Wu et al., 2019). Physiological response metrics have been well-studied to measure cognitive load (Charles & Nixon, 2019; Dias et al., 2018). These physiological metrics have potential for capturing some of the cognitive constructs

of NTS such as decision-making and situation awareness. For example, measures of brain activity through electroencephalography (Jacobs et al., 2006) and eye movements from eye-tracking technology (de Winter et al., 2019) have been used to investigate operator performance in a simulated task environment. Behavior is described as actions individuals conduct during an interaction with others or the environment. Metrics of these behaviors have also been used to measure NTS interpersonal skills such as teamwork and communication. Measures of behavioral actions are communication centered: studies have focused on speech and language analysis through discourse analysis and classification of spoken dialogue (Lingard et al., 2004; Tiferes et al., 2015). These objective metrics observed in other domains may address the limitations of subjective measures used in current behavioral checklists by raters, particularly of cognitive skills that may not be observed by raters. Therefore, objective metrics has the potential for continuous measurement to be used for monitoring the surgical team.

To the best of our knowledge, there has been no work critically examining and summarizing objective metrics of NTS in surgery. A review is warranted to guide integration of quantitative measures into current observer-based NTS assessments. The aim of this scoping review is to 1) identify current objective measures of NTS constructs in surgery, 2) evaluate the methodological rigor of current measures through a critical appraisal, and 3) discuss potential applications of these evaluations to take initial steps for objective NTS assessment of surgeons.

3.2 Methods

3.2.1 Search Terms Selection & Article Identification

A scoping review was completed following the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (Figure 1; PRISMA-Scr) guidelines (Tricco et al., 2018). After consultation of a Library Sciences faculty member, five databases were selected and searched that encompassed the clinical, behavioral science, and engineering literature: PUBMed, PsycINFO, Compendex, Inspec, and Scopus. Free text and controlled vocabulary (e.g., MeSH term, Subject, Thesaurus) were selected after review of widely-cited NTS literature (Catchpole et al., 2007; Undre et al., 2007; S. Yule et al., 2008). The search terms were categorized into four concepts: NTS, objective metrics, setting, and population (Table 3.1). The free text terms were used for every database. Control terms for each respective free text term were identified for

all databases except in Scopus, which does not include controlled terms. For example, search of the free text term "heart rate variability" yielded "heart rate" and "heart rate determination" in the controlled term database in PUBMed (i.e., MeSH). Both controlled terms were included in the final search.

The search terms were iteratively selected and tested to ensure all necessary terms were included. Overall terms describing each concept were included (e.g., NTS, behavior, physiological). For NTS, construct terms were added to the search. Objective metrics included common physiological and behavioral metrics used to measure NTS constructs (e.g., verbal and non-verbal variables during face-to-face interaction) (Charles & Nixon, 2019; Dias et al., 2018; Kazi et al., 2019). To narrow the scope of the search results, the setting was focused on the OR and the population was only surgeons. Common alternative spellings of terms were included, e.g., operating theatre for OR. The final search queries can be found in the Appendix (Table A1).

Table 3.1 Common free text search terms used in all databases (full search queries included in Appendix Table A1)

Concept	**Free Text Search Terms**
NTS	non-technical; non-technical skills; nontechnical; human factor; human factors; communication; teamwork; team work; leadership; situation awareness; situational awareness; vigilance; monitoring; decision making; decision-making
Objective Metrics	behavioral; behavior; behavior; assess; evaluation; objective; measure; empirical; quantitative; speech; interaction; gesture; movement; physiological; heart rate; heart rate variability; HRV; ECG; EKG; electrocardiography; skin conductance; skin conductance level; SCL; electrodermal activity; EDA; galvanic skin response; GSR; blood pressure; ocular; eye-tracking; eye tracking; brain measure; brain activity; EEG; electroencephalography
Setting	surgery; surgical; operating; operation; operating room; operating rooms; operating theatre; operating theatres
Population	clinician; surgeon

3.2.2 Article Screening and Selection

The final search was completed on October 14, 2019. Articles were downloaded systematically, and EndNote X9 was used to identify and remove duplicates. Rayyan QCRI (Qatar Computing Research Institute) was used for the title and abstract screening. Articles were included if they met the inclusion and exclusion criteria described in Table 3.2.

Table 3.2. Inclusion/Exclusion Criteria

	Criteria
Included	• Study is peer-reviewed literature • Study must be written in the English language • Study must include description of NTS assessment tool or non-technical skills (i.e., overall or specific construct) measurement/assessment e.g., Non-technical skills constructs include communication, teamwork, leadership, situation awareness, decision making • Study must include measurement of NTS/NTS construct that are objective (e.g., not only checklist- or survey-based) • Study population includes surgeons • Study setting is during an operation (intraoperative) or simulated operating room setting
Excluded	• Studies that are publication types of biography, historical article, duplicate publication, review article, or annals

Two researchers independently reviewed the titles and abstracts against the inclusion and exclusion criteria. All articles were reviewed twice by the author and her advisor (DY) reviewed a random 10% of articles following previous published protocol (McMullan et al., 2020). Full-text articles were downloaded and assessed for eligibility if the inclusion of an article could not be determined from the initial review. Articles to be included in the final synthesis were reviewed and agreed upon by the author and her advisor (DY).

3.2.3 Quality Assessment

Both researchers independently performed a critical appraisal of all included articles using the Mixed Methods Appraisal Tool (MMAT version 2018; Hong et al., 2018; Pluye, Gagnon, Griffiths, & Johnson-Lafleur, 2009). Each study was categorized into a study category (i.e., qualitative, quantitative randomized, quantitative non-randomized, quantitative descriptive, or mixed methods). Two screening questions were completed for all categories, and the studies were

evaluated as “Yes”, “No”, or “Can’t tell” to their respective methodological quality criteria. The “Can’t tell” response was used for a criterion if the information was not reported or not clear. Any discrepancies were discussed, and consensus was reached for all ratings.

3.3 Results

3.3.1 Study Selection

Figure 3.1 shows the PRISMA flow diagram summarizing the procedure and the articles included in each step. The initial download from the five databases resulted in 19,682 articles and 16,320 articles remained after duplicate removal in EndNote. Rayyan identified an additional 879 duplicates, which resulted in 15,943 articles included in the title and abstract review. JC performed this review of all articles, and DY reviewed 10% (1,598) subset of the articles. A total of 315 articles were retrieved for full-text review by JC, which included 21 articles identified by DY. Twenty-three articles were finally identified to meet the inclusion criteria by both researchers.

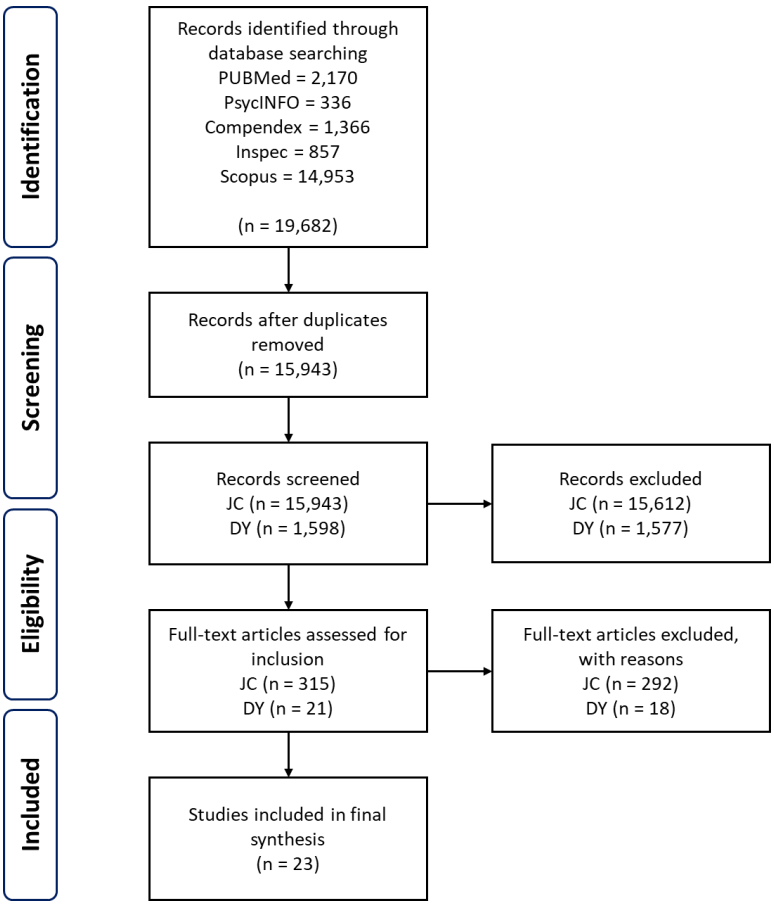

Figure 3.1. PRISMA diagram summarizing search strategy and study selection

3.3.2 Study Characteristics

Of the 23 articles, 16 were observational studies while seven were experimental studies. Six studies were completed in a simulation setting while 17 were within the OR (i.e., during an operation). Table 3.3 summarizes characteristics of each article as follows: description of the setting (OR/simulation); the research aim or hypothesis investigated; study design (experimental/observational); reported sample size, and key findings.

The objective metrics used to measure the NTS construct or multiple constructs reported in each study are shown in Table 3.4. Also included is the decomposition of each objective metric into specific dimensions, or quantified variables, that were reported in each article. For example, the frequency of communication events was identified as an objective metric. This frequency was extracted to different dimensions, or variables, such as counts of responses and questions or counts of the number of times one individual spoke to another (i.e., information flow). Nineteen articles

included metrics associated with interpersonal skills of communication, teamwork, or coordination. Six studies focused on cognitive skills of situation awareness (n = 4) and decision making (n = 2). The results from Table 3.4 are consolidated in Table 3.5, where the objective metrics are grouped to the frequency of communication metrics, other reported frequency metrics, and physiological metrics. The frequency of communication identifying topic/purpose (n = 14) and information flow (n = 11) were the most common variables quantified. Only five studies reported physiological metrics, which focused on brain activation and gaze metrics. Figure 3.2 summarizes the intersections of objective metrics with NTS within the clinical environment of surgery.

Table 3.3. Study details from the included articles (n = 23)

Lead author (year)	Setting	Research Aim/ Question/ Hypothesis	Study Design	Sample Size	Findings
Bezemer (2016)	OR	Gain insight in the involvement of non-operating surgeons in intraoperative surgical decision making	Observational	11 laparoscopic cholecystectomies	• 2 component features of decision making for clipping of cystic duct (decision point): participation and rationalization • Participation (e.g., degree of agreement sought prior to decision point) was found in 7/11 cases by consultant surgeon • Rationalization (e.g., verbal explication of evidential grounds justifying clipping) occurred in 9/11 cases
Cheriyan (2016)	Sim	Determine intraoperative noise levels during percutaneous nephrolithotomy and effects of noise on surgical team communication	Experimental	4 physicians	• Greater than 95% correct responses for first assistant, anesthesiologist, and circulator during ambient noise level condition • Correct response rate decreased for conditions with equipment noise and simulated noise (i.e., music) for each team member • Noise level of simulated surgical environment was comparable to a passing freight train at 30 feet
Cunningham (2012)	Sim	Investigate the design of a spatial aid for a collaborative surgical task. Authors hypothesized that presence of a spatial communication aid would improve performance time, reduce volume of communication, and improve efficiency of communication for novices.	Experimental	29 subjects	• Performance time was faster with cardinal direction aid than grid and control • Significant difference of volume of communication among camera degree conditions but not spatial aids • Lower ratio of communication in cardinal directions and grid conditions compared to no aid, representing increased degree of collaboration

Table 3.3 continued

Keller (2016)	OR	Test whether 1) Noise impairs case-relevant communication during surgeries (i.e., surgical teams engage in less case-relevant communication during high noise) and 2) Noise peaks impair case-relevant communication more when junior surgeons are in charge than senior surgeons	Observational	109 operations	• Noise peaks are associated with less case-relevant communication • Case-relevant communications decreased under high noise peak conditions when junior surgeons were in charge but not when senior surgeons were in charge • Case-irrelevant communication did not decrease under high noise level conditions
Korkiakangas (2014)	OR	Explore what factors affect slow or fast transfer of objects between surgical nurses and surgeons	Observational	20 operations	• 2 factors that affect object transfer were instrument trolley position and alignment • Instrument trolley position and alignment affects communication with surgeon and consequently speed of object transfer • Object transfer was faster when scrub nurse was standing close to surgeon and "converged" to follow the surgeon's movements than when the scrub nurse did not follow the surgeon's movements
Leff (2017)	Sim	Investigate differences in quality, confidence, and consistency of surgical decision making using functional neuroimaging	Experimental	22 subjects (10 medical student novices; 7 residents; 5 attending surgeons)	• Novices showed significant activation of dorsolateral prefrontal cortex during unprimed decision-making condition • No statistical significant activation of all expertise groups during primed condition • Residents and attending surgeons were significantly more certain and decision quality was superior than novices

Table 3.3 continued

Lingard (2004)	OR	Systematically describe the content and effects of case relevant communication events, and define and classify common communication failures	Observational	28 surgical procedures	• 31% of communication events were categorized as communication failures • Failure types were categorized into occasion (46%), content (36%), purpose (24%), and audience (21%) • 36% of communication failures visibly affected system process (e.g., efficiency, resource waste, delay)
Moss (2004)	OR	Evaluate a methodology for determining information needs of a data collection tool; document communication patterns of a OR charge nurse with developed tool; and characterize information needs for OR coordination	Observational	4 OR suites in 3 tertiary hospitals	• Most frequent communication target of charge nurse was OR nurse (39%); significant association between purpose of communication and communication target • Most frequent communication modes were face-to-face (69%), telephone (18%), and intercom (7%) • Significant difference between duration of communication episode and purpose of communication
Nyssen (2010)	OR	Investigate how robotic surgery induces changes in collective work using communication as a sign of the adaption process	Observational	Study 1: 9 cases Study 2: 36 cases Study 3: 2 cases	• Average case duration of robotic surgery was longer than laparoscopic case • Number of communication acts was reduced with repeated experience; number of communication acts regarding orientation, manipulation, and strategies was significantly reduced when both surgeon and trainee were experts of robotic system • Conversion from robotic to classic procedure is associated with an increase number of verbal communications

Table 3.3 continued

Raheem (2018)	OR	Understand different ways surgeons communication with bedside assistants during robot-assisted surgery and to identify the most efficient way of communication among surgical team members	Observational	26 robot-assisted surgeries	• 5 identified tasks: instrument change; clipping suction; irrigation; and retraction • Non-specific requests were most frequent during instrument change task; specific requests were most frequent for suction task; specific and non-specific requests were similar frequency for clipping task • Significantly shorter median action times for incomplete requests than complete requests
Santos (2012)	OR	Analyze and characterize cross-professional communication flow patterns in pediatric cardiac surgery	Observational	10 pediatric cardiac procedures	• Frequency of communication between pairs: main surgeon and scrub nurse (16%), main surgeon and first surgical assistant (13.8%), and main surgeon and perfusionist (12.4%) • Types of communication was varied between different roles: main surgeon to scrub nurse comprised of 84.2% request, main surgeon to first surgical assistant was 59.9% statements, and perfusionist to main surgeon was 65.4% answers • Communication was closed-loop between main surgeon and perfusionist but mostly open among other team members
Sevdalis (2007)	OR	Investigate case-irrelevant communication in the operating room	Observational	48 general surgery procedures	• Irrelevant comments and queries accounts for 50% of case-irrelevant communication • Surgeons initiated 33% and received 66% of case-irrelevant communication • External staff initiated most distracting communication

Table 3.3 continued

Sevdalis (2012)	OR	Map the types and initiators of communication in elective open versus laparoscopic surgery, their purpose, and their content	Observational	20 open and 20 laparoscopic inguinal hernia repairs	• No significant difference in mean operative duration, communication frequency, and communication rate between open and laparoscopic procedures • Communication was 80-81% initiated by surgeons and received by another surgeon (46-50%) or nurse (38-40%) • Communication in laparoscopic cases were significantly more related to equipment, providing direction, and consulting than in open cases
Sexton (2018)	OR	Investigate the impact of anticipation as a measure of efficiency in robot-assisted surgery	Observational	12 robot-assisted radical prostatectomies	• 31% of requests were anticipated; anticipation negatively correlated with operative time • Team familiarity negatively correlated with inconveniences • Significant correlation of surgeons' cognitive load with anticipation ratio, percent nonverbal requests, and total request duration
Thomas (2019)	Sim	Compare the efficacy of communication from robot assisted surgery (Da Vinci Si) speaker system to a wireless, hands-free audio system	Experimental	4 members of surgical team	• Accuracy of communication was increased with wireless, hands-free system than conventional robotic system • Significantly fewer correct phases when using conventional system for bedside assistant, anesthesiologist, and circulating nurse • No significant difference in number of correct phrases between different team roles when using wireless system

Table 3.3 continued

Tien (2010)	Sim	Use eye-tracking information of surgeons as a probe to measure situation awareness during simulated laparoscopic cholecystectomy	Experimental	8 novices and 8 experts	• Experts tended to glance at vitals screen more often than novices for both patient conditions • Two novices looked at secondary monitor only during higher-risk patient condition • Novices reported higher workload scores for both patient conditions than experts
Tien (2011)	Sim	Examine the relationship between vigilance and surgical skills, and show that novices and experts have different eye gaze patterns during a simulated laparoscopic procedure	Experimental	16 surgeons and medical students	• For stable patient, novices spent approximately the same mean duration of time looking at anesthesia monitor (0.9 s) as experts (0.8 s) • For unstable patient, novices spent less time looking at anesthesia monitor (1.6 s) than experts (3.2 s) • For unstable patient, only 3 novices checked vitals screen while 5 experts checked vitals screen
Tiferes (2016)	OR	Evaluate design and feasibility of data collection methods to capture and assess team activity during robot-assisted surgery	Observational	37 robot-assisted procedures	• Characterize communication into flow, mode, topic, and form • Identification of physical movement of personnel (e.g., most movement occurred between Circulating Nurse Zone and Transit Zone 1) • Classification of procedural interruptions (39% interruptions were related to surgical procedure)
Tiferes (2016)	OR	Understanding the nature of multimodal interactions between surgeons and bed side assistants	Observational	6 robot-assisted radical prostatectomies	• Identify 6 most frequent interaction topics: suction (22%), wash (18%), hold (11%), clip (11%), catheter (9%), switching/needle (8%) • Significant relationship between interaction type (verbal/nonverbal) and topic between surgeon and bed side assistant • Suction, wash, and hold interactions required minimal verbalizations

Table 3.3 continued

Tiferes (2019)	OR	Characterize verbal/nonverbal interactions among console surgeon, physician assistant, and scrub nurse to understand communication during robot-assisted surgery	Observational	11 robot-assisted radical prostatectomies	• Percentage of nonverbal interactions differed significantly by pair (e.g., 66% for Surgeon-Physician Assistant, 25% for Surgeon-Scrub Nurse) • Significant dependence between topic and percentage of verbal and nonverbal events for all pairs • Significant association between familiarity level and median percentage of verbal events
Wadhera (2010)	OR	Measure cognitive demands among OR staff, identify critical events, and develop and implement a protocol-based communication tool	Observational	18 cardiovascular surgeries	• Significant decrease in frequency of communication breakdown per case after communication protocol implementation • Decrease in frequency a call-back between surgeon and perfusionist post-protocol implementation • Nonverbalized critical actions per case decreased after protocol implementation
Weigl (2018)	OR	Identify type and severity of surgical flow disruptions and its influence on perceived intraoperative teamwork	Observational	40 robot assisted radical prostatectomies	• Highest flow disruption occurred during robot docking phase • Most severe disruptions related to communication and coordination during prerobot and docking phase • Significant relationship between disruptions and perceived intraoperative teamwork among surgeons

Table 3.3 continued

Zheng (2011)	OR	Investigate if vigilance – measured through eye-tracking techniques – is a function of surgeon experience in performing a laparoscopic procedure and whether a surgeon's vigilance was affected by simulated patient conditions (i.e., stable and unstable)	Experimental	23 surgeons	• Expert surgeons scanned patient vital signs (saccade eye movements) more often than novice surgeons • Experts increased frequency of checking anesthetic monitor for unstable patient (from 2.5 times for stable patient to 2.9 times for unstable patient) • Novices increased scan frequency from 1.1 times with stable patient to 2.1 times with unstable patient

Abbreviations: OR = operating room; sim = simulation.

Table 3.4. NTS constructs and objective metrics reported in included studies (n = 23)

Lead author (year)	NTS Construct Evaluated	Objective Metric	Specific Dimension Reported

Bezemer (2016)	Decision-making	Frequency of decision-making dimensions (i.e., participation and rationalization) and degree of agreement (i.e., unilateral/multilateral and implicit/explicit)	Decision making component features
			• Participation: degree to which agreement was sought prior to decision point
			  ○ Unilateral: no agreement was sought
			  ○ Multilateral: comments were made that explicitly designed to invite others to participate in decision making
			• Rationalization: verbal explication of visual evidence for justifying a clipping decision
			  ○ Implicit: where evidential grounds were not verbally said
			  ○ Explicit: where decisions were verbally described

Cheriyan (2016)	Communication	Frequency of correct phases identified; noise level in decibels	• Percentage of correct responses • Sound levels in dBA
Cunningham (2012)	Communication	Frequency of communication metrics	• Communication volume: total number of communications per trial • Communication ratio: # instructor communications/ # task communications
Keller (2016)	Communication	Frequency of case-relevant communication and noise level peaks in decibels	• Case-relevant communication event: uninterrupted communication related to current patient/ procedure (e.g., patient-relevant communication, teaching, instructions) • Case-irrelevant communication event: communication unrelated to patient/ procedure (e.g., patient-irrelevant or humor) • Noise peak: recorded any noise level reached 70 dB(A) or higher

Table 3.4 continued

Korkiakangas (2014)	Communication/ Teamwork/ Situation awareness	Frequency of communication events; speed of instrument passing	Interactional event from each team member
			• Request, utterance, question, repetition, response
			• Response was vocal or non-vocal
			• Associated bodily conduct
			Object passing
			• Average speed of passing per case
			• Interaction involved
			  o Surgeon's signaling of a request for passing (i.e., vocal/ non-vocal)
			  o Scrub nurse's focus of attention at the time of the signaling (i.e., physical orientation of scrub nurse)
			  o Scrub nurse's response to the request for assistance
			• Context of item passing
			  o Participants (scrub nurse, surgeons)
			  o Objects (e.g., syringe, instrument, swab)
			  o Spatial arrangement (i.e., layout)
			Interactional arrangements of scrub nurse
			• Alignment with surgeon and operating field: converged to gaze at surgeon and operating field
			• Alignment with other people, objects and actions: converged gaze to other concerns

Leff (2017)	Decision making	Hemoglobin concentration from Optical Topography	• Changes in cortical oxygenated hemoglobin and deoxygenated hemoglobin from 22 channels
Lingard (2004)	Communication	Frequency of communication events	Communication event/ communication failure • Audience • Purpose • Occasion

Table 3.4 continue

Moss (2004)	Communication/ Coordination	Frequency of communication events	Communication episode • Purpose (e.g., schedule surgery, coordinate staffing) • Mode (e.g., face to face, telephone) • Target individual • Duration
Nyssen (2010)	Communication	Frequency of communication events	Ratio of communication acts/duration of surgery *100 • Verbal demands regarding orientation and localization of organs • Verbal demands regarding manipulation of instruments and/or organs • Explicit clarification of strategies, plans, and procedures • Orders regarding to tasks (e.g., cutting, changing instruments, cleaning camera) • Explicit confirmation of detection or action • Other communications referring to state of stress or relaxation
Raheem (2018)	Communication	Frequency of communication events	Tasks • Amount of information included within request (i.e., specific/ non-specific/ unclear) • Frequency of utilization • Time to execute task • Inconveniences (e.g., repeated requests, required further clarification, resulted in frustration) • Acknowledgements (i.e., request was verbally acknowledged by bedside assistant)

Table 3.4 continued

Santos (2012)	Communication	Frequency of communication events	Characterization of communication flow patterns • Frequency • Direction • Type (i.e., request, question, answer, statement, information, and explanation) • Content (e.g., instrument request) • Pattern (e.g., closed-loop communication) Factors including communication • Disturbing elements • Interdependency with other non-technical social skills (e.g., teamwork and leadership)
Sevdalis (2007)	Communication	Frequency of case-irrelevant communication events	Case-irrelevant communication event • Source (who initiated the communication) • Recipient (whom the source addressed) • Content/category of event (e.g., comment queries, patient-related, teaching, equipment/ provisions)
Sevdalis (2012)	Communication	Frequency of communication events	Communication • Initiator • Recipient • Content (procedure, anatomy, equipment, patient, unrelated) • Purpose (directive, informative, consultative, education/ teaching) • Type (primary tasks, OR environment, OR management/ coordination)

Table 3.4 continued

Sexton (2018)	Teamwork	Frequency of communication events; anticipation ratio	• Requests
			  ○ Personnel
			  ○ Equipment type
			  ○ Mode of communication
			• Anticipation ratio (i.e., ratio of total requests – inquired by surgeon – to anticipated requests – execution occurred only after inquiry)
			• Inconvenience index (i.e., sum of events such as communication breakdown, multiple requests for item, or item was not readily available)

Thomas (2019)	Communication	Frequency of correct phases identified	• Number of correct phrases recorded
Tien (2010)	Situation awareness/ Vigilance	Eye motion metrics (e.g., saccade movements, fixation duration)	• Fixation duration • Saccades (e.g., glances)
Tien (2011)	Situation awareness/ Vigilance	Eye motion metrics (e.g., fixation duration)	• Fixation duration • Saccades (e.g., glances)
Tiferes (2016)	Communication	Frequency of communication events	Team communication • Information flow (sender, receiver, time, and duration) • Mode (verbal/ nonverbal) • Topic • Statement function (verbal) • Form (nonverbal) Surgical workflow • Ambulatory pattern of each team member tracked • Density of movement between locations through link diagrams Procedural interruption event • Duration • Personnel involved • Cause • Mode of communication

Table 3.4 continued

Tiferes (2016)	Communication/ Teamwork	Frequency of communication events	Interaction event • Sender • Recipient • Type (verbal/ nonverbal) • Topic (e.g., bag, case-irrelevant, cut, patient condition/information, tool preparation/organization)
Tiferes (2019)	Communication	Frequency of communication events	Primary • Modality (verbal/ nonverbal) • Topic • Pair (sender and receiver) Secondary • Sequences (grouping single interaction events that share the same pair *and* topic and related to same task instance) o Percentage of verbal events per sequence o Verbal grounding criterion (number of acknowledgements and repetitions divided by the number of interaction events in a sequence) per sequence
Wadhera (2010)	Communication	Frequency of communication events	• Type of communication breakdown o Miscues o No call-back o Repeated communication exchange o Occurrence of nonverbalized critical action o Ambiguous or unstructured communication exchange • Occurrence of no communication or poor communication during critical phase
Weigl (2018)	Teamwork	Frequency of disruption types	• Surgical flow disruption (e.g., event that disrupted procedure progress) o Source of disruption o Severity of interference rating

Table 3.4 continued

Zheng (2011)	Situation awareness/ Vigilance	Eye motion metrics (e.g., saccade movements, fixation duration)	• Number of saccades to anesthetic monitor • Eye fixation (percentage of time) on anesthetic monitors

Table 3.5. Summary of objective metrics found in included articles

	Frequency of Decomposed Communication Metrics					Frequency of Non-Communication-Based Metrics			Physiological Metrics	
Lead Author (year)	**Information Flow**	**Topic/ Purpose**	**Case Relevance**	**Mode (Verbal/ Non-verbal)**	**Breakdown /Failure**	**Disruption/ Inconvenience/ Interruption**	**Agreement**	**Correct Phases**	**Brain Activation**	**Gaze Metrics**
Bezemer (2016)							X			
Cheriyan (2016)								X		
Cunningham (2012)	X									
Keller (2016)			X							
Korkiakangas (2014)	X	X		X						X
Leff (2017)									X	
Lingard (2004)		X			X					
Moss (2004)	X	X								
Nyssen (2010)	X	X		X						
Raheem (2018)		X				X				
Santos (2012)	X	X				X				
Sevdalis (2007)	X		X							
Sevdalis (2012)	X	X								
Sexton (2018)	X	X				X				
Thomas (2019)								X		

Table 3.5 continued

Lead Author (year)	Frequency of Decomposed Communication Metrics					Frequency of Non-Communication-Based Metrics			Physiological Metrics	
	Information Flow	**Topic/ Purpose**	**Case Relevance**	**Mode (Verbal/ Non-verbal)**	**Breakdown /Failure**	**Disruption/ Inconvenience/ Interruption**	**Agreement**	**Correct Phases**	**Brain Activation**	**Gaze Metrics**
Tien (2010)										X
Tien (2011)										X
Tiferes (2016)	X	X		X		X				
Tiferes (2016)	X	X		X						
Tiferes (2019)	X	X		X						
Wadhera (2010)		X			X					
Weigl (2018)		X	X			X				
Zheng (2011)										X

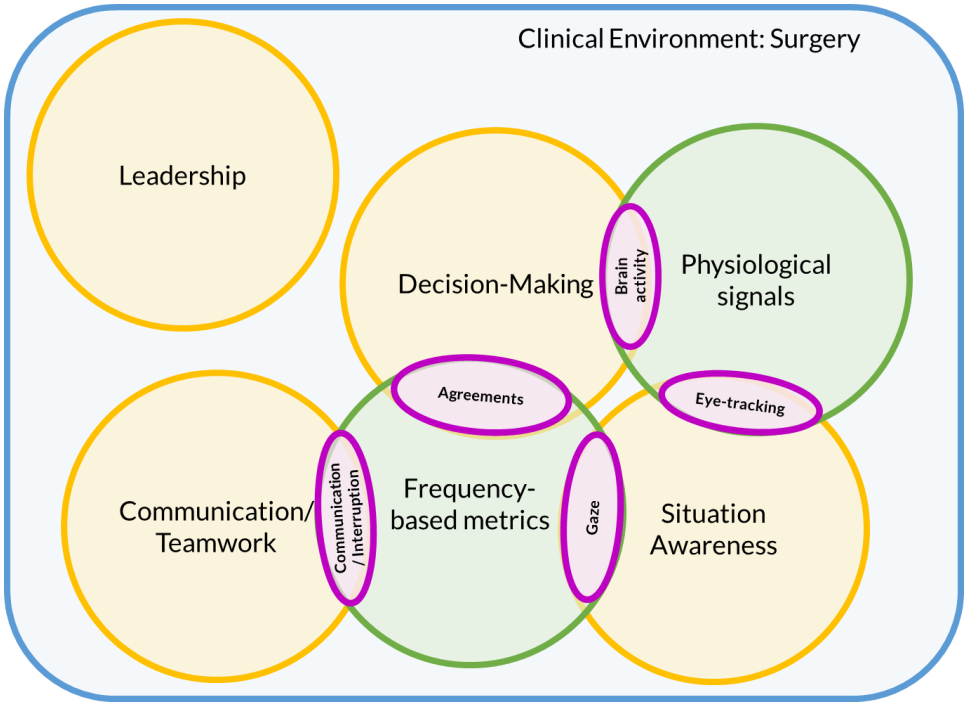

Figure 3.2. Visualization of the objective metrics (green) reported to measure NTS constructs (yellow) within surgery (blue), with metrics associated with NTS construct in the intersections (purple)

*Communication/ Interruption represents all communication-based frequencies and metrics such as flow disruption and inconveniences

3.3.3 Critical Appraisal

Table 3.6 summarizes the appraisal of the methodological quality of the included articles. The studies were categorized as a quantitative non-randomized study (n = 9), quantitative descriptive study (n = 12), or mixed methods study (n = 2). Research questions for four studies was unclear; consequently, it was unclear if the data collected answered the stated research questions for five studies. For the quantitative non-randomized studies (n = 4), there was ambiguity among the representativeness of the participants for the target population: Can't Tell (CT) was noted for these studies that did not note sampling strategy. In addition, confounding variables were not accounted for or noted in one study, and it was unclear if they were considered in two studies. A majority of the quantitative descriptive studies met the MMAT criteria; however, the sampling strategy was noted as CT if it was not described in the methods (n = 3). CT was also assigned to the criteria of "is the risk of nonresponse bias low" if the studies did not describe their sampling or used convenience sampling, which may increase bias of those who participated (n = 5). One

study (Bezemer et al., 2016) included only one "Yes" rating of the seven criteria. Additionally, the research question as well as interpretation and rationale of reporting the study components was ambiguous for one mixed methods study (Korkiakangas et al., 2014).

Table 3.6. Critical appraisal of included articles using the MMAT criteria

	SCREENING QUESTIONS		**QUANTITATIVE NON-RANDOMIZED STUDIES**				
Lead author (year)	Are there clear research questions?	Do the collected data allow to address the research questions?	Are the participants representative of the target population?	Are measurements appropriate regarding both the outcome and intervention (or exposure)?	Are there complete outcome data?	Are the confounders accounted for in the design and analysis?	During the study period, is the intervention administered (or exposure occurred) as intended?
Cheriyan (2016)	YES	YES	CT	YES	YES	YES	YES
Cunningham (2012)	YES	YES	CT	YES	YES	YES	YES
Keller (2016)	YES	YES	YES	YES	YES	YES	YES
Leff (2017)	YES	YES	YES	YES	YES	YES	YES
Nyssen (2010)	YES	CT	CT	YES	YES	CT	CT
Thomas (2019)	YES	YES	YES	YES	YES	YES	YES
Tien (2010)	CT	CT	CT	YES	YES	NO	YES
Tien (2011)	CT	CT	YES	CT	YES	CT	YES
Zheng (2011)	YES	YES	YES	CT	YES	YES	YES

	SCREENING QUESTIONS		**QUANTITATIVE DESCRIPTIVE STUDIES**				
	Are there clear research questions?	Do the collected data allow to address the research questions?	Is the sampling strategy relevant to address the research question?	Is the sample representative of the target population?	Are the measurements appropriate?	Is the risk of nonresponse bias low?	Is the statistical analysis appropriate to answer the research question?
Bezemer (2016)	CT	CT	CT	YES	CT	CT	NO
Lingard (2004)	YES	YES	YES	YES	YES	YES	YES
Moss (2004)	YES	YES	YES	YES	YES	YES	YES
Raheem (2018)	YES	YES	CT	YES	YES	CT	YES
Santos (2012)	YES	YES	CT	YES	YES	CT	YES
Sevdalis (2007)	YES	YES	YES	YES	YES	YES	YES
Sevdalis (2012)	YES	YES	YES	YES	YES	YES	YES
Sexton (2018)	YES	YES	YES	YES	YES	YES	YES
Tiferes (2016)	YES	YES	YES	YES	YES	CT	YES
Tiferes (2016)	YES	YES	YES	YES	YES	CT	YES
Tiferes (2019)	YES	YES	YES	YES	YES	YES	YES
Weigl (2018)	YES	YES	YES	YES	YES	YES	YES

Table 3.6 continued

	SCREENING QUESTIONS		MIXED METHODS STUDIES				
	Are there clear research questions?	Do the collected data allow to address the research questions?	Is there an adequate rationale for using a mixed methods design to address the research question?	Are the different components of the study effectively integrated to answer the research question?	Are the outputs of the integration of qualitative and quantitative components adequately interpreted?	Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?	Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?
Korkiakangas (2014)	CT	CT	YES	YES	CT	YES	CT
Wadhera (2010)	YES	YES	YES	YES	YES	YES	YES

CT = Can't Tell

3.3.4 Objective NTS Metrics

Within the included studies, behavioral or physiological metrics were used to quantify NTS constructs or used these objective metrics as independent variables to investigate the outcome of experimental conditions. Decomposed communication metrics were quantified in 19 studies within the OR, while teamwork was also associated with surgical flow disruptions in one study (Table 3.3). Additionally, situation awareness was associated through gaze metrics, identified by rater annotations or eye-tracking technology in three studies. Optical topography metrics were used in one study to measure decision-making. Within the included articles, no objective metrics were reported for measuring the leadership construct.

Synthesizing the reviewed literature, a framework for integrating the identified objective metrics into surgeon NT skill assessment is proposed (Figure 3.3). Drawn from a simplified model of one-way communication, this framework integrates the NTS constructs with the objective metric dimensions reported in the literature (Flin & O'Connor, 2017). Information or the meaning a sender intends to transmit are encoded, transmitted, and decoded by a receiver. In this process, both the sender and receiver must use their cognitive skills (e.g., decision making or situation awareness) to gather information and interpret the message. Interpersonal skills (e.g., communication/ teamwork or leadership) will either aid or hinder the transfer of the message, and poor skills may interpret the successful transmission of information. These skills can be quantified with the objective metrics found in this review. Dimensions of these metrics are listed in the dotted boxes. The following sections describe the NTS constructs measured in relation to the objective metrics.

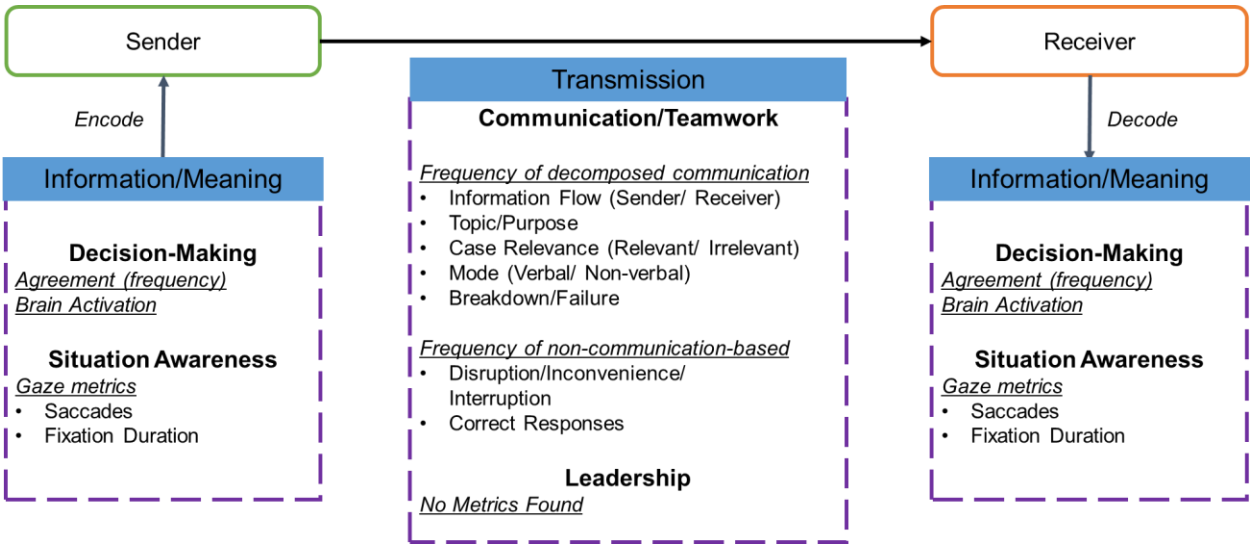

Figure 3.3. Model of one-way communication with reported NTS constructs (**bold**) and objective metrics (*italicized and underlined*) with specific dimensions (bulleted)

Communication and Teamwork Metrics

Studies investigating communication, teamwork, or coordination focused on quantifying communication, or interaction, events as outcomes. These three constructs are often jointly considered in the NTS assessment tools, thus was aggregated in this review. The interaction events were defined as a verbal or non-verbal exchange between two or more persons (Lingard et al., 2004). Studies utilized frameworks from linguistics to develop or implement communication coding schemes for communication patterns. Specifically, factors such as the number of communication acts, their purpose (e.g., content-based classification such as requests or questions), the type (verbal or non-verbal), and persons involved were identified. Twelve of the 18 articles reported these factors as frequencies that were used to describe characteristics of the surgery such as the type of procedure.

Four studies utilized communication metrics as dependent variables to evaluate different experimental conditions. Two studies investigated the influence of noise levels in an OR environment: Cheriyan et al. (2016) determined that an increased noise level (i.e., from ambient noise, equipment, and music) in a simulated OR led to decrease in correct response rate of surgical team members while Keller et al. (2016) quantified environment noise peaks in live ORs and found

that noise peaks are associated with less case-relevant conversations. The latter study also found that noise peaks occurred more often when a junior surgeon was in charge compared to when a senior surgeon was leading the operation. Furthermore, two studies investigated the effect of technology on communication: Thomas et al. (2019) examined the effect of different communication sources during robot-assisted surgery (i.e., from a traditional robotic console speaker or wireless headsets) and Cunningham et al. (2012) used communication ratio and volume to evaluate two different spatial communication aids to complete a simulated laparoscopic task. Variables such as the frequency of case-irrelevant communication and the number of errors were reported in three and two studies, respectively.

Additional metrics obtained from quantifying communication included variables that measured teamwork such as inconveniences and anticipation ratio. Inconveniences were measured by two studies as actions that required repeated requests or clarifications which influenced a communication breakdown (Raheem et al., 2018; Sexton et al., 2018). The two studies related the frequencies of communication event metrics as well as team familiarity and time to perform a task. Furthermore, anticipation ratio – the ratio of total requests to anticipated requests (i.e., requests executed after inquiry from surgeon) – was found to be negatively correlated with operative time, and 31% of total requests observed in the study was anticipated (Sexton et al., 2018). Finally, Weigl et al. (2018) associated self-reported teamwork with the frequency of flow disruptions, which included disruptions from communication, equipment, and procedure during robot-assisted surgeries.

While intersections between the construct and objective measure exists, directional relationships between the two were often not reported. Four studies utilized communication metrics as dependent variables to evaluate different experimental conditions. Specifically, these studies reported variables such as correct response rate as a measure for quantifying noise in the OR environment or evaluating technology. Furthermore, perceived teamwork was correlated with metrics such as the number of inconveniences or flow disruptions; however, these associations between the frequencies of communication-based metrics were quantitative descriptive studies that aimed to describe feasibility and distributions of the metrics and not report causal relationships.

All the communication studies required trained observers to analyze the stated communication metrics; a range of 1 to 4 observers were reported for the observational studies, and one study did not report the number of raters. Additionally, only six studies reported a metric

of agreement or reliability (e.g., interrater or interobserver reliability) to classify the communication metrics. Studies reported a range of interrater agreement between 0.74 - 0.98 (Keller et al., 2016; Raheem et al., 2018; Sevdalis et al., 2012; Sexton et al., 2018; Tiferes, Bisantz, et al., 2016, p. 20; Tiferes et al., 2019), which is typically reported as substantial agreement (McHugh, 2012).

Decision-Making Metrics

Decision-making was objectively measured in two studies (Bezemer et al., 2016; Leff et al., 2017). Bezemer et al. (2016) used linguistic analysis to identify decision points during a procedure and to quantify the decision-making metrics of "participation" or "rationalization." Participation was measured as unilateral or multilateral agreement, and rationalization was classified as implicit or explicit. Additionally, another modality for inferring decision-making were physiological signals. Leff and colleagues (2017) utilized hemoglobin concentration from optical topography (OT) to measure surgical decision making of different expertise levels and experimental conditions (i.e., primed and unprimed conditions). Measuring changes in cortical oxygenated hemoglobin and deoxygenated hemoglobin from 22 channels from OT, the authors concluded that novices showed significant activation of the dorsolateral prefrontal cortex during unprimed surgical decision making (i.e., an experimental condition that did not show the next operative step) which was not observed in experts.

Situation Awareness Metrics

Four studies reported metrics of situation awareness: these studies measured gaze of participants to infer situation awareness or vigilance. Three studies used eye-tracking technology and were from the same research group (G. Tien et al., 2011, 2010; Bin Zheng et al., 2011); one used manual annotation of gaze (Korkiakangas et al., 2014). The studies that leveraged an eye-tracking device investigated the differences of eye motions such as the number of saccades (referred to as "glances") and fixation durations for novice and experts during experimental conditions that facilitated the need for an increased monitoring of a simulated patient monitor (i.e., stable and unstable patient conditions). Zheng et al. (2011), which included the greatest number of participants of the three studies, reported that expert surgeons showed greater saccades on the

patient monitor than novices and this frequency increased during the unstable patient condition than stable condition. Furthermore, Korkiakangas et al. (2014) measured the degree of situation awareness of surgical team members as the alignment of gaze, which was affected by the relative position of the personnel. The alignment of gaze from surgical team members were observed and noted by researchers. In all studies, the number of saccades and fixation time in respective to objects in the environment were metrics leveraged to infer situation awareness.

3.4 Discussion

The goal of this study was to identify objective metrics of NTS and discuss opportunities and limitations for their application for improving NTS assessment. From the scoping review of five clinical and engineering databases, 23 articles were found to meet all inclusion criteria for this search. The handful of identified literature shows that quantitative metrics are mostly frequency-based and there need to be more research on understanding and utilizing objective metrics of surgeon NTS.

NTS of surgeons was the primary focus of this study, as individuals in this role are typically responsible for high-risk tasks (i.e., surgeries) where errors and adverse events may lead to safety consequences for other team members and for patients. The significance of such leaders to understand and display good NTS enhances current surgical training by contributing to the continuous learning of surgeons themselves but to trainees and OR staff as well (Agha et al., 2015). Objectively measuring such skills for surgeons is the first step in obtaining deeper understanding of NTS in the OR, which can lead to future applications of evaluations of the additional team members and interventions to mitigate errors for improved training and patient safety.

3.4.1 NTS Measurement Through Linguistics

Though objective metrics were identified, 16 studies used observer-based methods based on a linguistic framework to quantify the NTS metrics. For example, the 12 studies identifying the communication events were annotated by raters either real-time or through video post-operatively. These studies often reported at least two raters; however, there is still a subjective bias with the identification of the manual coding. Similar to the metrics measuring interpersonal skills, Bezemer

et al. (2016) identified and classified decision-making points through linguistic analysis, which may have been biased due to the dependence of raters.

Specific dimensions reported in studies including the communication construct were often vague and varied. For example, content of a communication was decomposed and classified by studies in different ways: by purpose (e.g., questions or requests) or by its relevance to performing the surgical task (e.g., case relevant or irrelevant). Recognizing the different communication taxonomy in the literature, Tiferes et al. (2015) completed a systematic review to develop a standardized coding scheme of OR communication to minimize the variance. The review proposed names for communication event dimensions such as information flow for sender/receiver and topic/theme to describe the content of the communication. These metrics were identified as dimensions for the objective metric of communication frequency in this scoping review (Tiferes et al., 2015). It should be noted that seven articles included in this scoping review was included in the literature used to build the taxonomy (Cunningham et al., 2012; Lingard et al., 2004; Moss & Xiao, 2004; Nyssen & Blavier, 2010; Santos et al., 2012; Sevdalis et al., 2007, 2012; Wadhera et al., 2010). The popularity of using linguistic analysis for NTS rating supports the importance that communication has on all NTS constructs.

The use of objective communication features in the majority of the studies indicates its fundamental role in the identification and assessment of NTS constructs. As the NTS framework for the surgical team was developed, principles of NTS for assessing pilots' Crew-Resource Management (CRM) skills were applied, which notes that skills that are evaluated should be those that are directly observable or inferred from monitoring communication or other behaviors (Flin et al., 2003; S. Yule, Flin, Paterson-Brown, et al., 2006). Hence, in the NOTECHS system for CRM, it is stated that the communication construct is not distinguished separately since it is inherent in all behaviors underlying the other constructs (Flin et al., 2003). This supports the use of the communication metrics to assess NTS constructs objectively; however, there is a potential that physiological signals can be used to measure additional cognitive skills and unobservable behaviors that were reported preliminary NTS taxonomy such as mental readiness and workload distribution (S. Yule, Flin, Paterson-Brown, et al., 2006).

3.4.2 NTS Measurement Through Physiological Metrics

Several studies used physiological metrics to measure the cognitive NTS constructs of decision-making. These studies used the metrics and relationships established in existing literature to associate the physiological response to NTS to investigate another study factor. For example, the changes of hemoglobin concentration from OT was summarized as activation of brain regions and was used to detect surgical decision-making by experts and novices (Leff et al., 2017). Conclusions inferring decision-making were drawn based on previous studies identifying activation of the dorsolateral prefrontal cortex (DLPFC) during deductive reasoning and decision requiring working memory and effort (Dixon & Christoff, 2014; Owen et al., 2005). Furthermore, there is a capability of utilizing brain activity to directly measure cognitive control and decision-making; previous literature has reported relationships between brain activity in the DLPFC from functional magnetic resonance imaging (fMRI) with attentional control (MacDonald et al., 2000; Milham et al., 2001; Philiastides & Sajda, 2007) and good choice behavior in a behavioral decision-making task (Yarkoni et al., 2005). It should be noted, however, that though readings from physiological sensors are not influenced by observers, the application of this technology may not be feasible during a live operation due to the obtrusive nature of the technology; thus, assessment of decision-making during surgery is still limited.

Eye-tracking metrics were associated with situation awareness in three studies. The study team for these studies inferred situation awareness and vigilance as the physiological signals for their study design, citing the established literature on shifts of eye gaze and attention (Deubel & Schneider, 1996; Just & Carpenter, 1975). Moreover, the authors used metrics obtained to quantify eye-behavior in the environment (e.g., number of saccades) to differentiate experts and novices (G. Tien et al., 2011, 2010; Bin Zheng et al., 2011), but there is a potential to further use this technology. Eye-tracking equipment not only can map gaze on the environment, but it can measure pupillary response. This physiological response has been utilized in the healthcare domain to measure task difficulty (Wu et al., 2019) and surgical skill (Richstone et al., 2010). Leveraging these metrics may provide additional insight into not only situation awareness, but also additional cognitive skills such as decision making or workload distribution. For its application in the OR, mobile eye-tracking technology can be used with minimal intrusion to determine an individual's perception of the environment but may be limited in measuring their comprehension or knowledge of the environment.

3.4.3 Methodological Rigor and Limitations of Studies

From the completed critical appraisal, the methodological rigor of the studies varied. Though 11 studies were rated "Yes" for all criteria, five studies included "Can't tell" responses for at least one of the two screening questions, which may indicate that the paper is not an empirical study (Hong et al., 2018). These papers were often brief in the purpose of the study and methods followed to address the research aim or question. For quantitative descriptive studies, the authors agreed upon the dependence of the risk of nonresponse bias with sampling strategy: the bias was rated unclear if the sampling strategy was not explicitly stated. The varied ratings of the studies may indicate the need for more rigorous study designs and analyses to objectively measure NTS in surgery.

Several of the included studies were conducted by the same research group (i.e., overlapping authors) and used subsets of data from a greater dataset. Five studies investigating communication from the same research group were included in this review (Raheem et al., 2018; Sexton et al., 2018; Tiferes, Bisantz, et al., 2016; Tiferes et al., 2019; Tiferes, Hussein, et al., 2016): these articles helped create an understanding the communication requirements and patterns during robotic-assisted surgery. Communication was also investigated during open and laparoscopic procedures in work by Sevdalis and colleagues (2007, 2012), who used frequency of communication events to quantify case-irrelevant conversations and overall communication in the OR. Furthermore, another study team inferred situation awareness and vigilance using eye-tracking metrics in three studies, with the results from each study building upon the previous. Finally, though the two observational studies by Bezemer et al. (2016) and Korkiakangas et al. (2014) utilized video analysis in the OR, different NTS constructs were investigated in each study. The publications by these study teams show each of their unique focus on improving surgeon performance utilizing NTS constructs; however, additional work is needed to specifically measure these interpersonal and cognitive skills.

There were also gaps found in the existing studies to establish relationships of the metrics and current NTS constructs and tools. The purpose of 13 studies were to describe or characterize factors of surgery while two were exploratory. The objective metrics of NTS (e.g., counts of communication events) were used as dependent variables to quantify factors that influenced surgery such as flow disruptions or inefficiencies or test independent factors such as communication modes or differences between expertise levels. Interestingly, studies that reported

dimensions of communication as the dependent variable suggests that the independent variable investigated (e.g., environment noise levels) may be covariates that influence NTS of the surgical team. With the nature of these studies, it is difficult to draw clear conclusions regarding the relationship between the metrics and NTS. Moreover, there was an absence of objective metrics linking the NTS construct of leadership. Studies that reported on leadership that were not included in the study due to the absence of objective metrics focused on assessing different leadership styles and were questionnaire-based (Barling et al., 2018; Hu et al., 2016). This suggests that leadership elements (e.g., maintaining standards and coping with pressure) may be difficult to measure with current objective measures. The limited hypothesis-testing studies and lack of clear relationships or patterns reveals the limited use of objective metrics for NTS.

Furthermore, a critical gap of the identified literature is investigating the current standardized method of NTS assessments through joint use of ratings and objective metrics. No studies identified reported on using the stated behavioral or physiological measures to relate with NTS performance, and only one study related a communication metric (i.e., flow disruption) with self-perceived teamwork (Weigl et al., 2018). Articles that used the standard assessment tools were removed during the screening process as they did not include an objective metric and were often focused on evaluating NTS of cohorts or training programs (Wood et al., 2017).

3.4.4 Limitations of Methods

Less than 1% of articles that were identified from the initial search were included in this final study. Due to the broad search terms that were used to capture as many objective metrics (e.g., assess or evaluation), many studies that did not meet the scope of this search were identified. Additionally, there was a high exclusion of articles after the full-text review (>85%). These excluded articles often discussed NTS constructs but did not have an objective measurement technique. For example, articles of NTS measurement using the subjective measures were common and excluded due to the lack of objective measures. Furthermore, several articles on decision-making between patient-clinician for surgery were identified in clinical databases while articles on human-machine interaction and teams resulted from engineering journals. These articles were excluded for this review, but this shows the broad use of the selected search terms in the literature that may suggest a need for specific nomenclature across disciplines for this topic.

Limitations in the scoping review included the scope and screening. The scope of the population was narrowed to include only surgeons and no other surgical team members (e.g., nurses, trainees). As NTS literature primarily focuses on surgeons, this population was chosen for this review (Gordon et al., 2012; Leuschner et al., 2018); however, there has been an increase in the focus of additional team members in relatively recent years (Boet et al., 2018; Gordon et al., 2019; Gostlow et al., 2017). Additionally, simulation was not included explicitly as a search term, but studies conducted in simulation were not excluded; and only studies written in the English-language were included. Finally, the majority of article screening was completed by the author. Though the screening was completed twice, articles may have been missed. Future work should address these limitations by expanding the search scope and utilizing multiple screeners to completely review all article to gather a comprehensive understanding of objective NTS metrics in surgery.

This scoping review identified objective metrics of NTS constructs and discussed applications of their use in a surgical environment. Behavioral and physiological features were identified in 23 studies to measure communication, teamwork, decision making, and situation awareness. The use of these metrics to quantify NTS in surgery can be used to minimize the bias present in current assessment metrics that are check-list based of observed behaviors. This review reveals the current gap in the literature for objective NTS measurements in surgery, and further work should investigate additional metrics and technologies that measure behavior and physiology to measure NTS.

4. OBJECTIVE NON-TECHNICAL SKILLS MEASUREMENT IN ROBOTIC-ASSISTED SURGERY

4.1 Introduction

Non-technical skills (NTS) are cognitive (e.g., decision-making and situation awareness), social (e.g., teamwork and leadership), and personal resource (e.g., cognitive workload and stress management) skills that are technically relevant for safe and efficient task performance (Flin & O'Connor, 2017). Specifically, these skills are interconnected: management of cognitive load and stress during a task can reduce an individual's capacity to respond to distractions and slow decision making (Speier et al., 1999). These skills are especially relevant in surgery, where adverse events have detrimental effects on surgical outcomes (Arora et al., 2010; W. O. Cooper et al., 2019; Mazzocco et al., 2009). Poor NTS behaviors increase chances for errors and adverse events; thus, identifying, training, and assessing NTS can help improve intraoperative safety and efficiency (Flin & O'Connor, 2017; Siu et al., 2016).

Current methods for NTS assessment in surgery rely on checklist-based behavior rating systems. Evaluation tools are typically composed of several intrapersonal and cognitive NTS constructs, and scores for the individual constructs are often used to determine an overall NTS score. Each rating system is anchored by poor and good NTS behaviors: higher scores are given for exemplar behaviors that promote patient safety and lower scores are given for poor behaviors that are detrimental to safety. A recent systematic review identified 31 observational tools that quantified NTS for individuals and teams in the operating room (OR) (McMullan et al., 2020). This review concluded that the Non-technical Skills for Surgeons (NOTSS) has the strongest evidence of validity and reliability for NTS assessment of individuals; however, observational tools are limited in that they are affect by a rater's bias and are time intensive. Therefore, there is a need for objective measurements to minimize the subjectivity of these rater-based assessment tools.

Objective NTS measurements include metrics that can be quantified based on physiological response or behavior. The scoping review presented in Chapter 3 identified frequency-based metrics and physiological features that objectively measure NTS of surgeons. Dimensions of the frequency of communication metrics such as information flow (i.e., identification of speakers) and topic of communication were found to be associated with a surgeon's NTS. Moreover, behaviors

not identified in the scoping review, but has the potential to measure NTS have been found to be correlated with changes in workload and flow in the healthcare environment (Rosen et al., 2018; Tiferes, Hussein, et al., 2016). These include speech and location-based metrics such as prosodic elements of speech (e.g., pitch and intensity) and movement of OR staff. Of the studies identified in the scoping review, 35% of the studies involved NTS assessment with robotic-assisted surgery (RAS) technology.

The increased complexity of systems with RAS technology imposes unique challenges for NTS in the OR (Tiferes et al., 2015; Weigl et al., 2018). During a RAS procedure, the surgeon sits on a console away from the patient bed and surgical team to manipulate end-effectors that control the robotic arms, which are inserted through small incisions on the patient (Catchpole et al., 2019). Operating from the console decreases a surgeon's field of view – which may lower their situation awareness of the room – and often increases difficulty of communication, as the surgeon speaks into the console and a microphone is used to amplify their voice to communicate with the surgical team (Thomas et al., 2019). Due to these changes and challenges to surgical flow, team interaction during RAS has been explored, and methods to capture NTS with aspects of surgical performance and outcome have been identified (Ahmad et al., 2016; Tiferes et al., 2015, 2019; Tiferes, Hussein, et al., 2016).

Task performance metrics for intraoperative performance and skill include time and events that deter surgical team members from the task such as interruptions, distractions, or disruptions (Antoniadis et al., 2014; Koch et al., 2020; Weigl et al., 2018). Task performance metrics have been related to NTS, where poor NTS behaviors have been associated with avoidable incidences (Siu et al., 2016). As NTS are skills relevant to efficient task performance, these incidents, or events, that interrupts regular surgical flow and operative time can be used as objective metrics to infer NTS.

The purpose of this study was to develop objective models of surgeon NTS during RAS. To build a fully objective NTS model, behavior metrics and task performance metrics were used to predict NTS measured from the standard, rater-based assessment tools. Then, behavior metrics were associated with task performance metrics. We hypothesized that 1) objective behavior metrics can predict observational NTS scores, 2) objective task performance metrics can predict NTS scores, and 3) behavior metrics can be used to predict task performance metrics.

4.2 Methods

This study was approved by the University's Institutional Review Board. Thirty-four RAS cases were observed. The study population comprised of four male, right-hand dominant robotic surgeons. Procedures included those within general surgery and urology specialties.

4.2.1 Measurements and Equipment

NTS Assessment

The NOTSS tool was used by raters to assess surgeon NTS (S. Yule et al., 2008) which has been validated in the literature for its reliability and acceptability to measure surgeon NTS (Jung et al., 2018). This observational tool includes the NTS constructs of communication/teamwork, leadership, situation awareness, and decision making (Figure B1). Each construct is comprised of three elements, and a rater evaluates behaviors relevant to each element. The scores given to these elements are used to determine an overall construct score. The average of the constructs scores are used to calculate an overall NTS score. The rating scale for this instrument is centralized around patient safety, where the score of 1 represents behavior that may endanger patient safety and 4 represents behavior that enhances patient safety and can be used as models for others.

Self-Perceived Workload

Self-perceived workload was measured through the NASA- and SURG- Task Load Index (TLX) (Hart & Staveland, 1988; Wilson et al., 2011). Completion of both surveys yields nine unique domains that may influence the surgeon's workload (Figure B2). Both tools include the domains of mental, physical, and temporal demands. Three unique domains are included for each instrument: the NASA-TLX evaluates performance, effort, and frustration while the SURG-TLX evaluates task complexity, situational stress, and distractions. The scale anchors are comprised of adaptations of "very low" and "very high", with performance anchors noted as "perfect" and "failure" and situational stress marked as "not very anxious" and "very anxious". For all domains, anchors were numerically represented between 0 to 10. The unweighted overall NASA-TLX score

was calculated from a summation of the six included domains (Hart, 2006), and the SURG-TLX score was calculated by the same procedure.

Communication Metrics

Communication metrics for this study were adapted from a previously established coding scheme of team interaction in surgery (Tiferes et al., 2015). Communication was classified into metrics that included information flow, period, communication type/topic, and components of closed-loop communication. Information flow was determined by the different combinations of senders and receivers of communication. Since this work evaluates surgeons' NTS, only communication involving the surgeons were noted, i.e., the surgeon was included as a sender or receiver. Furthermore, the period was classified into the different phases of the surgical procedure, as explained in the following section. Communication type/topic was categorized into five topics that were adapted from Parush et al. (2011) and Hazlehurst et al. (2007). Table 4.1 includes a description and example of each topic. Finally, each communication was decomposed into components of closed-loop communication: the sender transmitting a message is classified as a call-out; the receiver acknowledging the message is classified as a check-back; and the sender confirming the correct interpretation or decoding of the message by the receiver is classified as a closed-loop (Bowers et al., 1998; Härgestam et al., 2013). Non-verbal communication, or behaviors to complete a task through actions, were also quantified. All communication metrics were quantified through video analysis by one rater who had previous experience annotating communication.

Table 4.1. Categorization of communication types/topics, adapted from Parush et al. (2011) and Hazlehurst et al. (2007)

Topic	Description	Example
Request	Requesting, directing, or instructing an individual to complete an action	"Pass me the suture."
Confirmation	Verifying or confirming an action was acted upon or to a statement	"Yes, that is correct."
Question	Asking an individual about a value, state, or action Note: if a statement was asking for a confirmation (e.g., "is this the correct patient?), it was classified as a question.	"Do you see any bleeding?"
Goal-sharing/Status	Sharing information to create understanding of current state or expectation of future state	"We're ready to dock the robot."
Case-irrelevant	Pertaining to non-case relevant topics (e.g., about another procedure or patient, non-work related)	"Is the next patient ready?"

Speech Metrics

Prosodic elements of speech have been associated to changes in individual emotional and cognitive stress as well as clinical performance (Mendoza & Carballo, 1998; Peng et al., 2019). Speech metrics such as fundamental frequency and amplitude have been associated with workload: fundamental frequency is positively correlated with cognitive workload in simulation and speech volume (i.e., intensity) increased with task workload at an *in situ* nursing station (Mendoza & Carballo, 1998; Rosen et al., 2018). Furthermore, features of duration, intensity, pitch, and rate have been previously associated with medical students' clinical performance, and a positive relationship between speech duration and performance score was also found (Peng et al., 2019). Definitions of the speech metrics are summarized in Table 4.2. Each metric was obtained during a speaker turn, which is the period of time a speaker is speaking (i.e., segment of communication). In this study, the surgeon's audio was obtained through recordings from a lapel microphone (Zoom H1, Zoom, Inc, Hauppauge, NY) attached to a voice recorder (RØDE smartLav+ Microphone, RØDE Microphones, Silver Water, NSW, Australia).

Table 4.2. Definitions of speech features

Speech Features	Definition
Speech Duration (s)	Amount of time surgeon spoke
Speech Pitch (Hz)	Relative low or high tone perceived by the ear
Speech Intensity (dB)	Perceived loudness
Speech Rate (1/s)	Total number of syllables/second

Proximity Metrics

Location or position tracking of surgical team members have been utilized in the literature to investigate surgical flow and relationships with procedure interruptions (Bayramzadeh et al., 2018; Tiferes, Hussein, et al., 2016). A study by Tiferes et al. (2016) characterized the movements of the surgical team through observations, and spaghetti diagrams were created to understand the density of movement of the OR staff during RAS procedures. Zones, or areas of the OR, were established in this study, and the frequency of OR personnel passing each zone was quantified. For this study, the physical location was summarized into metrics measuring distance relative to the surgeon. Specifically, locations of each team member were categorized into if he/she was close (<1 m), near (1 – 3m), or far (>3 m) away from the surgeon. The ultra-wideband-technology based Pozyx positioning system (Pozyx NV, Belgian) was used to obtain positioning data. The Pozyx anchors were placed in the corners of the OR, and tags with battery packs were placed in the surgical team members' pockets (Figure 4.1).

Figure 4.1. Pozyx tag placed in OR personnel's pockets for location positioning

Task Performance Metrics

NTS are technically relevant to the safety and for efficient completion of a task (Flin & O'Connor, 2017). In order to obtain method triangulation (Carter et al., 2014; Denzin, 1978) for further validity of the NTS measures, reference-standard metrics that are relevant to task performance were investigated in this study. Specifically, the task completion time and number of intraoperative incidents were quantified. Time to complete a task is often used as a technical performance metric to assess skill proficiency (Stefanidis, 2010), and it was used as a metric to infer OR efficiency in this study. Additionally, incidences were quantified as a performance metric. Intraoperative incidents include avoidable or unavoidable events that can lead to errors and decreased patient safety (Catchpole et al., 2007; Siu et al., 2016). A recent systematic review found that flow disruptions, or incidences that cause deviations in the operative procedure, are associated with increased OR time and are negatively associated with surgical outcomes (Koch et al., 2020). Increased incidences have been related to poor NTS, and inverse relationships between the number of interruptions and NTS scores have been reported (Gillespie et al., 2017; Siu et al., 2016). The number of incidents were annotated by observers through video analysis in this study and followed a previously published categories for incidences (Siu et al., 2016). Problems or events identified included minor or operative problems such as equipment failure or delay, breakdowns in communication or coordination, or distractions. These task performance metrics provide a metric for triangulation, supporting that behavioral data collected model NTS.

4.2.2 Data Collection

All RAS procedures performed by the participants were targeted and observed based on the surgeon's approval and researchers' availability and were completed using either the DaVinci Si or Xi model (Intuitive Surgical, Inc., Sunnyvale, CA, USA). All members of the surgical team were provided a study information sheet to inform them of the specific of the study. If a surgical team member did not wish to participate, researchers made efforts to not capture the individuals on the recording devices (e.g., move video camera so that he/she would not be filmed) and those individuals did not wear the proximity sensor. Any inadvertent recordings of those individuals were removed post-processing.

Each case was recorded after patients were draped and to the end of the procedure. A Go-Pro (HERO7, San Mateo, CA, U.S.A.) camera was used to record the OR. The attending surgeon wore a headset microphone, and the recorded video and the surgeon's audio were synchronized for further analysis. The NASA/SURG-TLX was completed by the participant at the end of each procedure.

Procedures were decomposed to five different surgical periods, or phases. The phase included robot docking; 5-minutes before critical; critical; 5-minutes after critical; and 10 minutes before robot undocking. The docking portion included the time when the robotic arms were moved from its resting location to above the patient until the surgeon (or assistant) started performing on the robot console. The critical phase was identified for each procedure by clinical subject matter experts. This included the nerve sparing portion of a prostatectomy and dissection near the femoral nerve, artery, and vein during inguinal hernia repairs. Ten-minutes before the undocking of the robotic arms was chosen as the other surgical team members often removes the robotic arms from the patient, and not the surgeon. An overview of the data collected and metrics obtained are summarized in Figure 4.2.

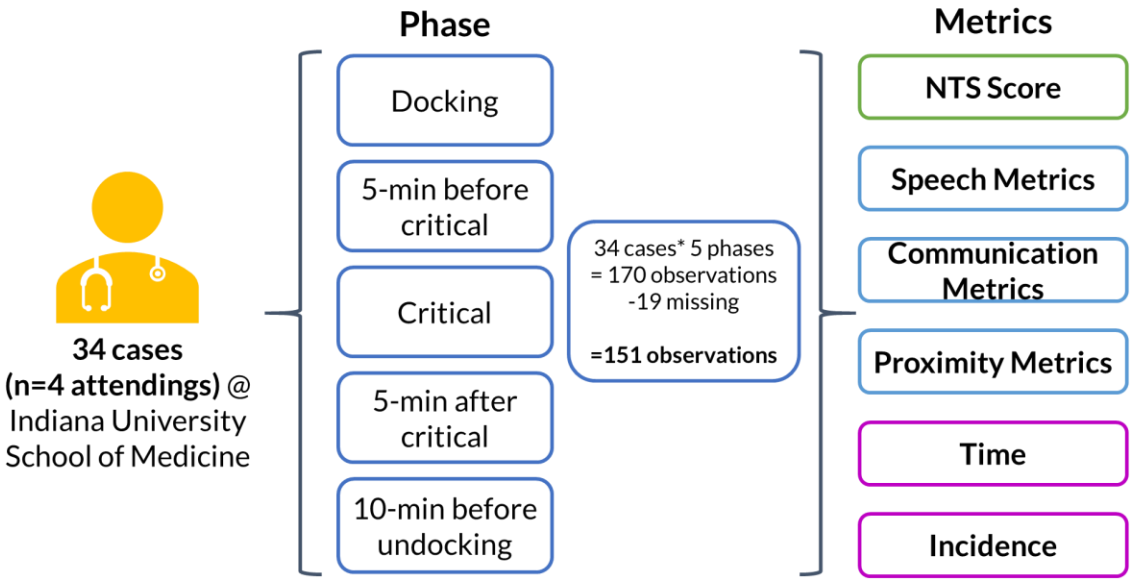

Figure 4.2. Overview of the decomposition of cases and metrics

4.2.3 Data Processing and Analysis

Video recordings and audio from the surgeon were synchronized. This synchronized video was used for video analysis. NTS assessments were completed for the participants during each phase by an expert rater. Any surgical team members who did that wish to participate in the study but were caught on either the video or audio were removed from the recordings manually.

Communication Metrics

A previously developed Excel Macro (Figure 4.3) was adapted to annotate the communications involving the surgeon (Yu et al., 2014). The frequency of the decomposed communication was obtained, and a summary of the features obtained are summarized in Figure 4.4. For each information flow pair with the surgeon and a team member, nine features were obtained, i.e., if there were five team members, then 45 total features were obtained. Communication to the whole team was also quantified (i.e., the surgeon stated a general remark not specified to one receiver). If there were more than one individual for a specific role (e.g., circulating nurse), communication was quantified individually and summed for further analysis.

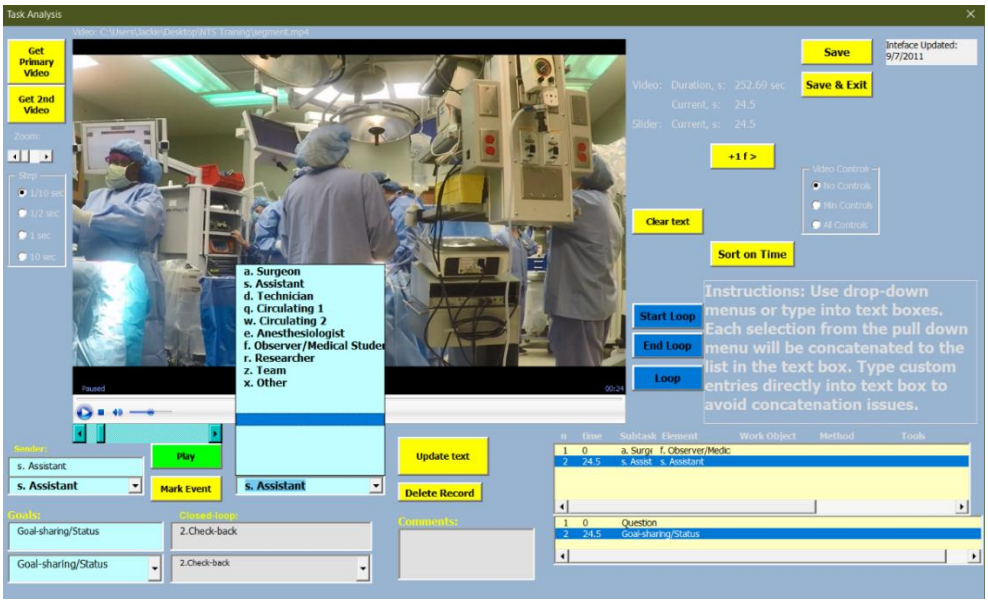

Figure 4.3. Excel Macro used to quantify communication metrics

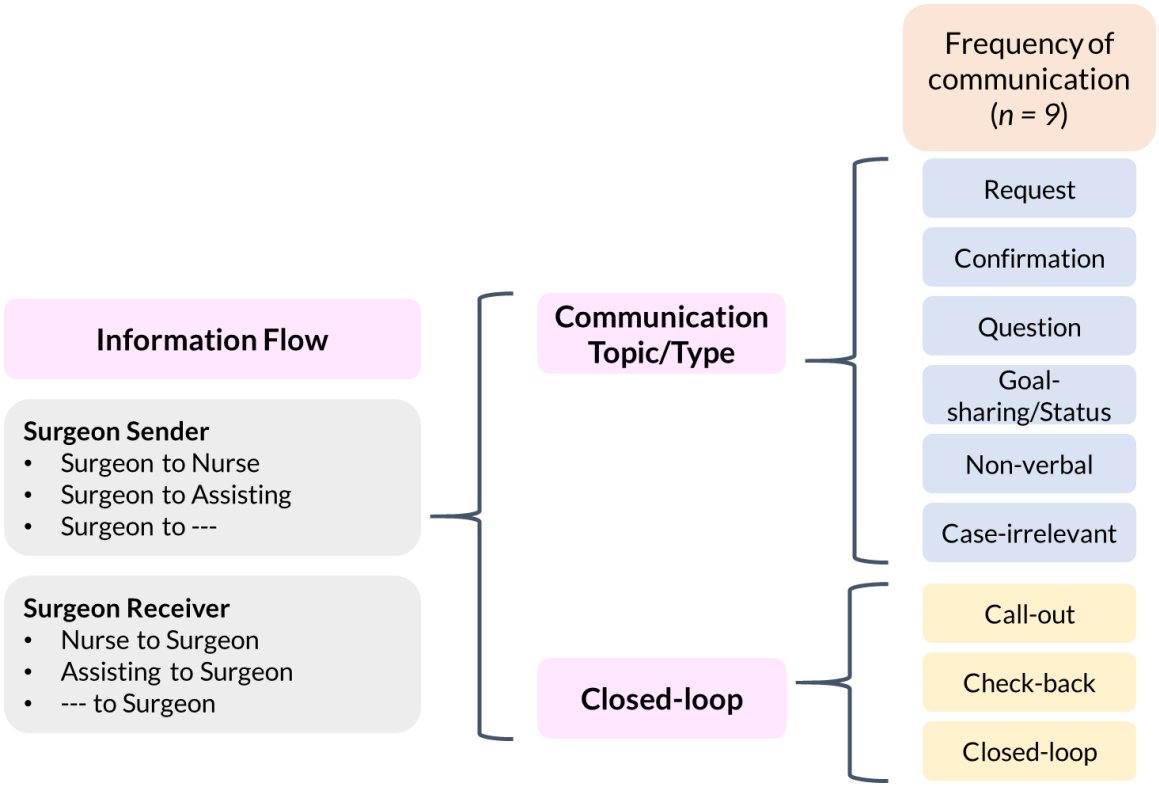

Figure 4.4. Summary of communication features

Speech Metrics

A previously established pipeline to extract prosodic elements of speech was followed (Peng et al., 2019). From the video recordings, all other environmental noise or non-surgeon voice were manually silenced. PRAAT (De Jong & Wempe, 2009) was used to process the surgeon's audio to extract the four speaker-turn based speech features: duration, intensity, pitch, and rate. Successive differences ($Feature_{Turnn} - Feature_{Turnn-1}$) were also calculated for each vocal feature to capture changes of the features over time. For example, these differences were calculated between Speaker Turn 2 and 1, as shown in Figure 4.5. Descriptive statistics obtained for these eight measures include minimum, maximum, mean, standard deviation, range, and interquartile range. In addition, the burstiness, or the temporal distribution of time spent speaking ($\sigma_{feature}/\mu_{feature}$) (Rosen et al., 2018) and root-mean-square-differences of each vocal feature during the speaker turn was calculated. Figure 4.6 summarizes the 56 speech features obtained.

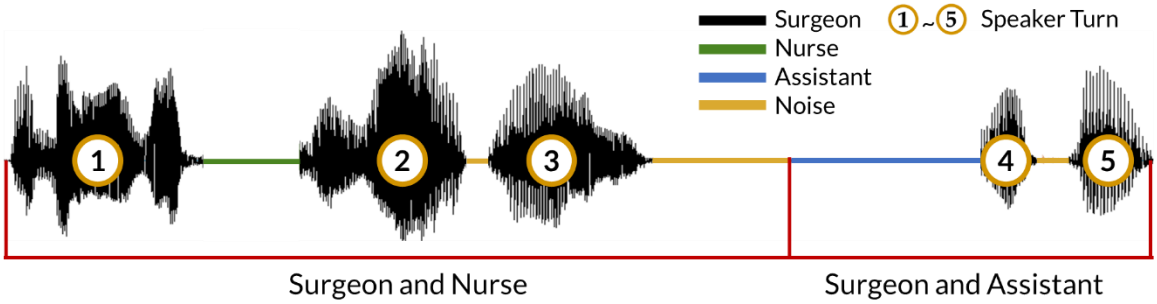

Figure 4.5. Sample audio of surgeon. The numbers represent an individual speaker turn

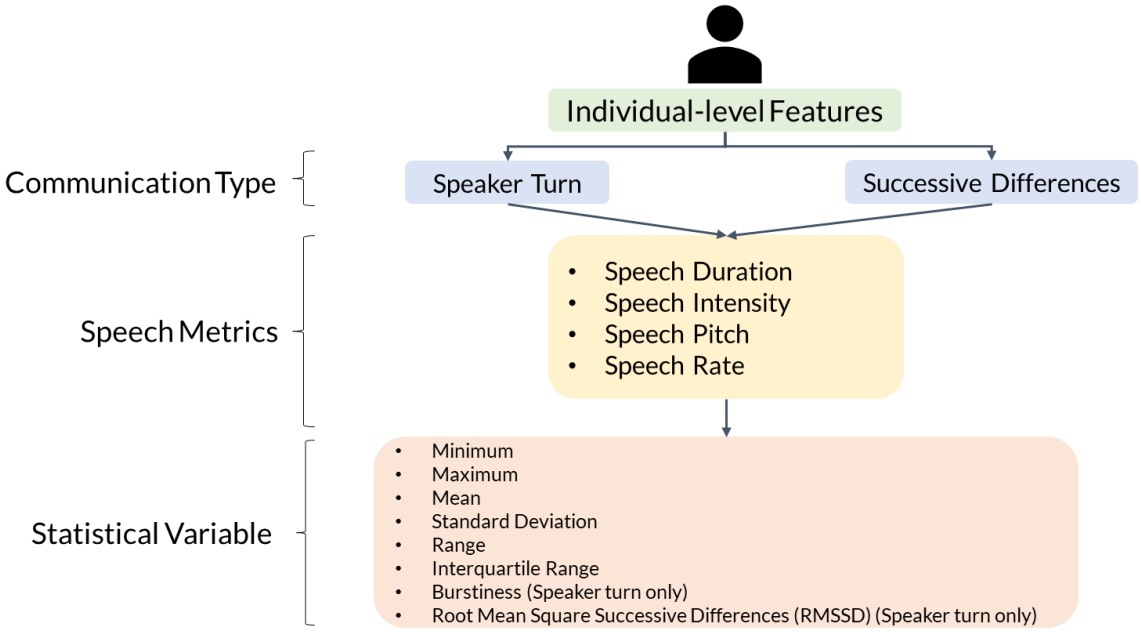

Figure 4.6. Summary of speech features

Proximity Metrics

Absolute positions within the OR and relative distance between a team member and surgeon was obtained. Proximity tags were also placed under the patient bed, robot arm base, and the DaVinci console. For each category of distance (i.e., close, near, and far), the percent of time the team member, or object, spent within that range of distance was calculated. For each of the three proximity bins with overall distance, descriptive statistics were calculated (Figure 4.7).

Nineteen features for each individual and object were obtained, i.e., if five team members other than the surgeon was present, then 95 features were calculated.

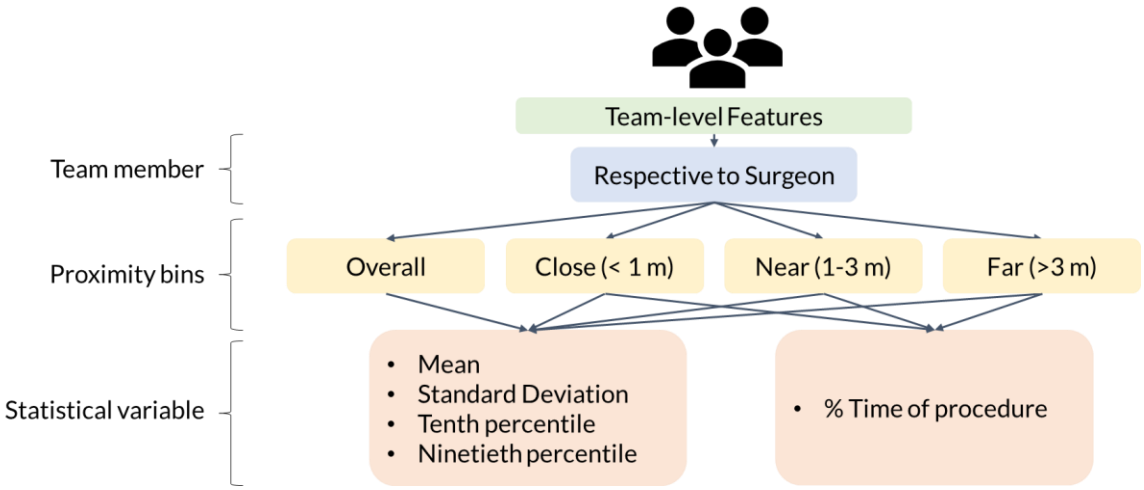

Figure 4.7. Summary of proximity features

Task Performance Metrics

Two task performance metrics were obtained for this study: duration of a surgical phase and the number of incidents. Among the five segmented procedure phases, the task performance metrics were obtained only for the docking phase. The before/after critical and undocking phases have fixed durations, and the critical phase was not selected due the time variability among different procedures and patients. Incidents that occurred in the whole operating room, not just of the surgeon, relating to poor behavior were quantified following the taxonomy in Siu et al. (2016). For the measurements during the docking phase, two raters assessed the surgeon NTS and annotated the number of incidents. One rater with a human factors background rated all NTS and quantified the number of incidents, and a second rater with a psychology background rated the docking phase subset of the data. Both raters had experience with NTS assessments, and a training session between the raters were completed with three cases to reach an agreement for the ratings. The average of the NTS scores and number of incidents between the raters were used for task performance analysis.

Data Pre-processing

The data used for this study was from an initial implementation of the sensors to obtain the behavioral metrics. Due to the preliminary implementation of the proximity sensor and equipment failure (e.g., lapel microphone stopped recording), missing data of the behavioral features needed to be considered. All existing data was used for the preliminary analysis prior to variable selection.

An overview of the feature reduction pipeline is shown in Figure 4.8. To initially extract variables, intra-correlated features were removed (e.g., a metric's mean, min, and max) and those using previously reported metrics in the literature were selected. Additionally, individual linear correlations with overall NTS score with each remaining feature was completed on the full dataset, and those with significant association ($p < 0.05$) were used as the objective behavior metrics. The selected features were then used for the subsequent analysis. Data was imputed using the Multivariate Imputation by Chained Equations (MICE) to estimate the any missing data (Buuren & Groothuis-Oudshoorn, 2010). Furthermore, due to a small distribution of NTS scores, classification of model and non-model NTS behavior was completed to gain insights on what features can be associated with the highest NTS ratings. Overall NTS score of 4 was assigned as model behavior and scores below 4 were categorized as non-model behavior. All data processing and statistical analysis was completed in R (Version 1.1.456, RStudio, Inc.).

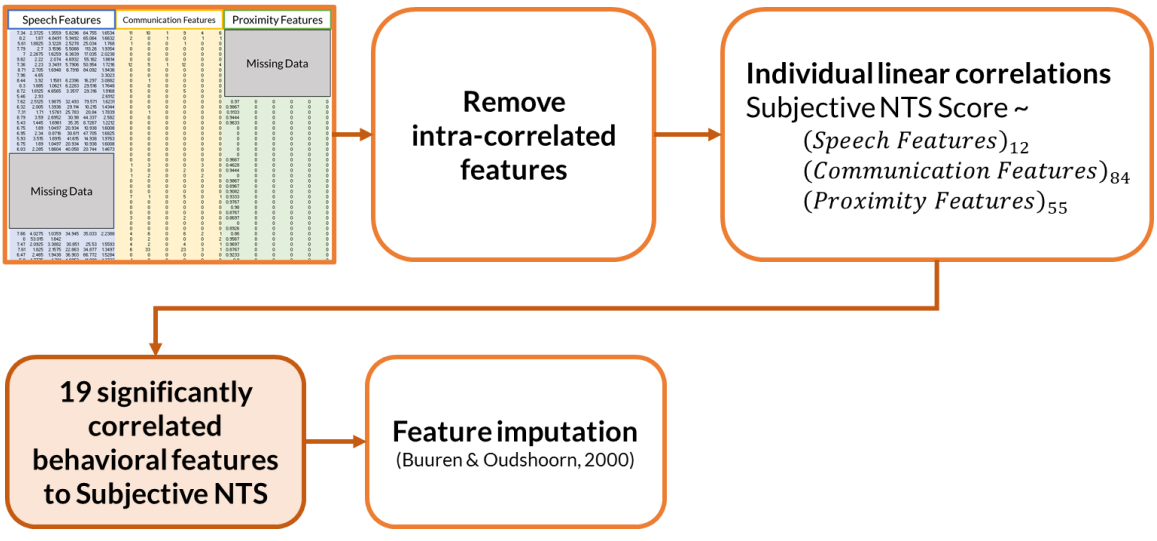

Figure 4.8. Overview of feature reduction pipeline

Statistical Analysis

To investigate the effect of potential confounders, such as the surgical phase and individual surgeon on overall NTS score, one-way analysis of variance (ANOVA) with a Bonferroni correction was performed for pairwise comparisons. NTS score was used as the response variable for separate ANOVA models for the phase and surgeon. Furthermore, Pearson's correlations were calculated with NTS score with each self-perceived workload dimension. Though the NTS score values are ordinal, since the average of the element scores were used to determine the construct score and the average of those construct scores were averaged for an overall NTS score, they were treated as continuous variables. In addition, since the workload survey was completed at the end of the procedure and was representative of the surgeon's self-perceived workload throughout the entire procedure, the average of the overall NTS scores determined for each surgical phase was used in the correlation calculation (i.e., the overall NTS score for the entire procedure was associated with perceived workload).

For predicting overall NTS scores, linear mixed effects and non-linear models were developed. The relationship between the behavioral metrics and NTS scores were analyzed with *lme4* (Bates et al., 2012): selected behavioral features were entered as fixed effects (without interaction terms), and surgeon were included as random effects. Predicted values from the developed model were used to compare with actual values. Machine learning algorithms were used to determine the accuracy of classifying model and non-model NTS behavior (*caret* package, Kuhn, 2008). One linear and three non-linear algorithms were used: Linear Discriminate Analysis (LDA), k-Nearest Neighbors (kNN), Support Vector Machine (SVM) with polynomial kernel, and Random Forest (RF). Three-fold cross-validation was performed to create a 70%-30% split of the training data set in each fold (Hastie et al., 2009). Twenty percent of the data was held-out and used as the testing set to determine the highest performing model, and a confusion matrix was created to determine the performance of the behavioral metrics to predict model NTS.

Separate linear models with the task performance metrics (i.e., duration and number of incidents) were developed to predict NTS score during the docking phase. A two-way intraclass correlation coefficient (ICC) was calculated for the overall NTS score and number of incidences to determine the agreement accuracy between the two raters. For a fully objective NTS model, mixed effects model to predict task performance metrics with behavioral metrics were used, and the model fit was evaluated by Likelihood Ratio Test against a null model.

4.3 Results

The number of cases observed, demographic, experience, and case load of each participants is shown in Table 4.3. Two surgeons had more than 20 years of experience while two surgeons had less than 5 years of experience. Distribution of overall NTS scores is shown in Figure 4.9: the minimum score was 2.92 and maximum score was 4, and the mean score across surgeons was 3.45 ± 0.43.

Table 4.3. Number of observations, demographic, experience, and case load of participants

Surgeon	n	Age	Years of Experience	MIS procedures (hours/week)	Robotic (hours/week)
Surgeon 1	8	62	36	10	10
Surgeon 2	13	48	20	14	4
Surgeon 3	1	38	4	15	4
Surgeon 4	12	31	1	15	6

n = number of case observations, MIS = Minimally Invasive Surgery.

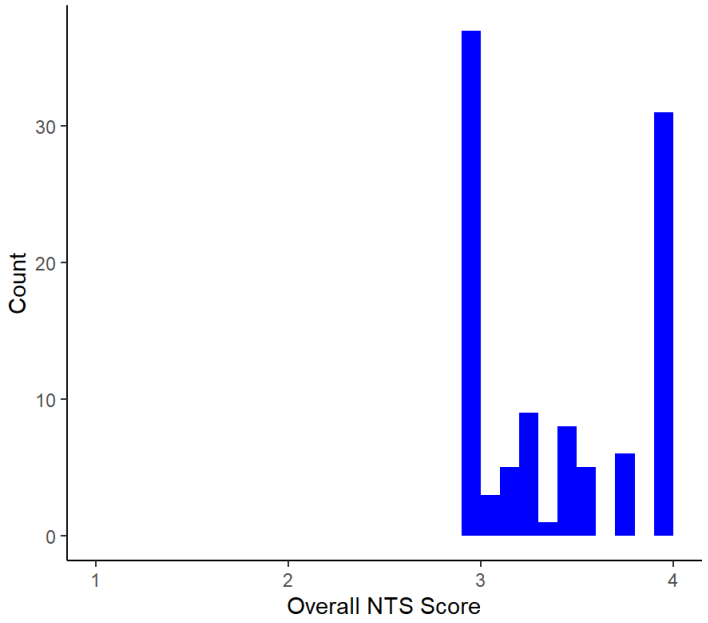

Figure 4.9. Histogram of overall NTS score (n=151) with 0.1 bin size

4.3.1 Effect of Surgeon and Phase Confounders on NTS Scores

Statistically significant differences of overall NTS score between surgeons were observed ($p < 0.05$), and score differences approached significance ($p = 0.06$) between the phases from the Kruskal-Wallis one-way ANOVA. Multiple comparisons of means with the Bonferroni correction (e.g., *p*-values were multiplied by the number of comparisons) showed significant differences between Surgeon 1 and all other surgeons ($p < 0.05$), with Surgeon 1 rated an average of 0.44 lower than the other three surgeons. No significant pairwise comparisons of phases were reported ($p > 0.05$).

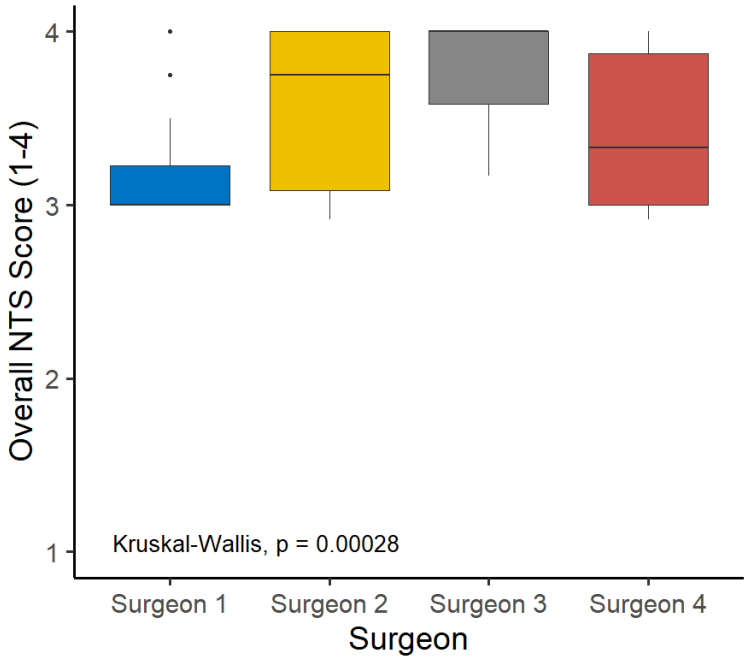

Figure 4.10. Comparison of overall NTS score by surgeon

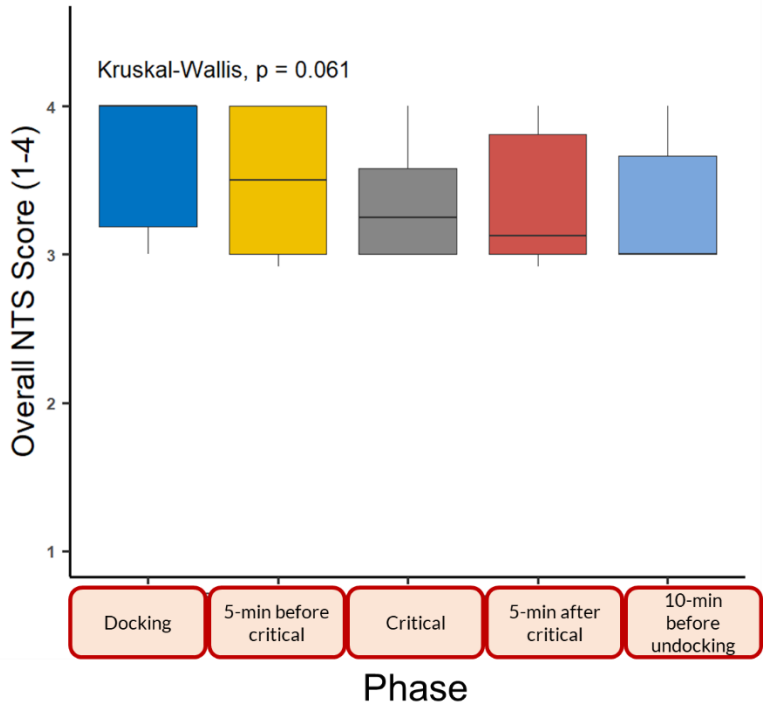

Figure 4.11. Comparison of overall NTS score by phase

4.3.2 Relationship between NTS Score and self-perceived workload

No significant correlations were found between overall NTS score and the self-perceived workload domains ($p > 0.05$). Though the relationships were not significant, moderate correlations ($0.67 > r > 0.39$) (Taylor, 1990) were found with the effort, frustration, task complexity, situational stress, and overall NASA-TLX domains (Table 4.4). Self-perceived task complexity was negatively correlated with NTS score while all other domains showed positive relationships.

Table 4.4. Correlations between overall NTS score with individual workload domains

Workload Domain	Pearson correlation coefficient (*r*)	*p* - value
Mental	0.13	0.70
Physical	0.19	0.57
Temporal	0.21	0.53
Effort	0.50	0.12
Performance	0.30	0.37
Frustration	0.50	0.12
Task Complexity	-0.38	0.25
Situational Stress	0.50	0.12
Distractions	0.35	0.29
NASA-TLX Overall	0.07	0.35
SURG-TLX Overall	0.25	0.46

4.3.3 Behavioral metric feature selection

Mean and burstiness values of the speech metrics were used for further analysis and the mean and standard deviation of proximity features were considered. Features with the DaVinci console and secondary persons for a specific role was not considered due to the presence of these individuals in only two procedures. From individual correlations between the remaining behavioral features and overall NTS score, 19 features remained: 12 communication features, 2 speech features, 5 proximity features. Table 4.5 shows the features that were significantly correlated with overall NTS score, and all individual correlations with each metric is in Appendix C. For surgeon-initiated communication, the data suggest that an increase in the frequency of callouts and requests communication with the circulating nurse increase with surgeon NTS score (r = 0.25-0.30, p < 0.02). Communication with the anesthesiologist such as callouts, questions, and requests were also significantly associated with NTS score (r = 0.21-0.27, p < 0.05). Additionally, it was found that non-verbal communication to the technician was significantly correlated (r = 0.22, p = 0.03). For surgeon-receiving communication, it was found interactions with the circulating nurse were significant in predicting NTS score: closed loop communication completed by the surgeon (r = 0.26, p = 0.02) and the number of questions that were asked to the surgeon were positively correlated (r = 0.28, p < 0.01). For speech features, mean speech duration (r = 0.43, p < 0.01) and pitch (r = 0.40, p < 0.01) of surgeons significantly predicted NTS score. Finally, features describing the percent time the assistant and circulating nurse spent in the different proximity ranges and one distance metric were significantly correlated with NTS score. Specifically, the

percent of time the circulating nurse far away from the surgeon had the strongest linear correlation with NTS score, with a moderate correlation coefficient of -0.45; this suggests the surgeon's NTS score decreased with longer time the circulating nurse spent more than 3 m away.

Table 4.5. Significantly correlated behavioral features with overall NTS score

Feature	**Pearson correlation coefficient (*r*)**	***p* - value**
Communication (*n* = 12)		
From Surgeon		
Circulating Callout	0.25	0.02
Circulating Request	0.30	<0.01
Circulating Confirmation	0.27	0.01
Circulating Non-verbal	0.26	0.01
Technician Non-verbal	0.22	0.03
Anesthesiologist Callout	0.22	0.03
Anesthesiologist Request	0.27	0.01
Anesthesiologist Question	0.21	0.05
To Surgeon		
Circulating Closed-Loop	0.26	0.01
Circulating Question	0.28	0.01
Technician Request	0.21	0.04
Anesthesiologist Confirmation	0.24	0.02
Speech (*n* = 2)		
Mean Duration	0.43	<0.01
Mean Pitch	0.40	<0.01
Proximity (*n* = 5)		
Assisting % Time Close	0.44	0.03
Assisting % Time Near	0.43	0.04
Circulating % Time Near	0.43	0.01
Circulating % Time Far	-0.45	0.01
Circulating Overall Mean Distance	-0.34	0.04

4.3.4 Prediction of NTS Score

With the 19 identified features from above, the missing data was imputed for the following analysis. A mixed-effects model was used to predict NTS score with objective behavioral metrics. Comparing the fitted NTS scores from the model with actual scores, there was a moderate correlation (r^2 = 0.33). From the visualization of the actual versus fitted graph (Figure 4.12), it can be observed that the model generally overestimated the low NTS scores while underestimated the

high (score = 4) NTS scores. Table 4.6 summarizes the fixed effects of the model: it was found that mean pitch and %time the assistant spent near the surgeon were significant predictors of NTS score ($p < 0.05$). To evaluate full model, the Likelihood Ratio Test was performed. The behavioral metric model was significantly different from the null model (i.e., model that estimates the mean of the data; $\chi^2(19) = 56.18$, $p < 0.01$).

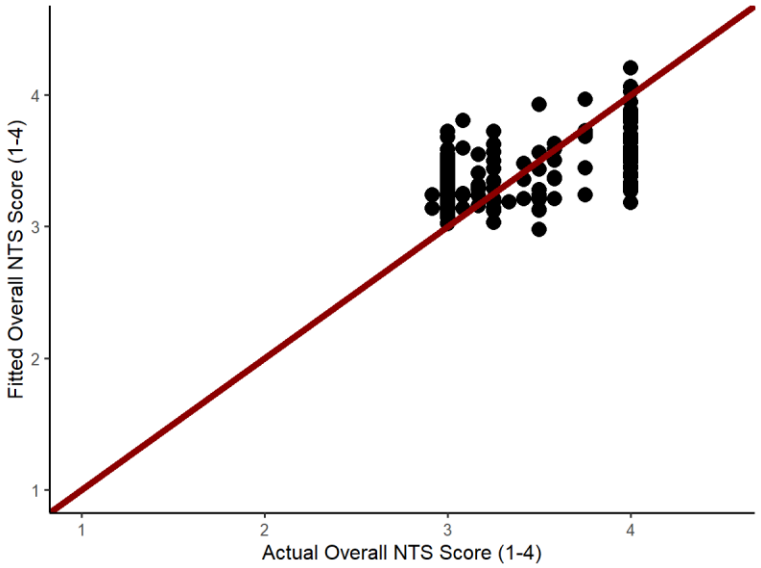

Figure 4.12. Actual v. fitted graph of the behavioral model predicting NTS

Table 4.6. Summary of mixed effects model of behavioral metrics on NTS score

Feature	**Estimate**	**Std. Error**	***p* - value**
Intercept	2.620	0.236	< 0.001
Surgeon-Circulating Callout	-0.009	0.027	0.749
Surgeon-Circulating Request	0.019	0.017	0.278
Surgeon-Circulating Confirmation	-0.070	0.088	0.429
Surgeon-Circulating Non-verbal	0.020	0.028	0.470
Surgeon-Technician Non-verbal	-0.001	0.012	0.948
Surgeon-Anesthesiologist Callout	0.061	0.062	0.327
Surgeon-Anesthesiologist Request	-0.120	0.108	0.267
Surgeon-Anesthesiologist Question	-0.040	0.160	0.804
Circulating-Surgeon Closed-Loop	0.084	0.150	0.574
Circulating-Surgeon Question	0.104	0.085	0.225
Technician-Surgeon Request	-0.015	0.112	0.894
Anesthesiologist-Surgeon Confirmation	-0.003	0.004	0.448
Speech Mean Duration	0.027	0.029	0.353
Speech Mean Pitch	0.005	0.002	0.004*
Assisting % Time Close	0.856	1.089	0.433
Assisting % Time Near	0.353	0.176	0.046*
Circulating % Time Near	-0.206	0.156	0.189
Circulating % Time Far	-0.221	0.192	0.253
Circulating Overall Mean Distance	-0.005	0.029	0.876

* indicates statistically significant ($p < 0.05$)

To classify model and non-model NTS scores, the 19 behavioral features were used as input to the classifiers. From 151 observations, 80% of the data was partitioned into a training set, which was further split into 3 sets with the size of 81, 81, and 80 for the 3-fold cross-validation. The summary the performance of the classification on the training set through the 95% confidence interval for the area under the receiver operating characteristic curve (AUC) of the one linear and three non-linear models are summarized in Figure 4.13. Though variability of the results is large as shown by the error bars, in general, the non-linear models appears to perform better than the linear model.

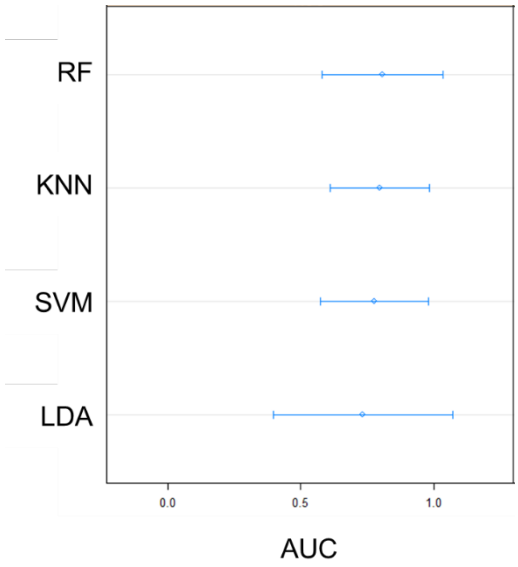

Figure 4.13. AUC of the three non-linear and linear models on the training set for model NTS classification

The random forest classifier achieved the best performance, and this model was validated on the test set. The confusion matrix of the results is shown in Table 4.7; the model achieved a classification accuracy of 0.77 (95% confidence interval [0.58, 0.90]) and F1 score of 0.82; however, the *p*-value for the Mcnemar's test was 0.45 for the testing classification. This indiiates that there is high variability for the prediction accuracy of the model.

Table 4.7. Confusion matrix of random forest model

		Actual class			
		<Model NTS	Model NTS		
Predicted class	<Model NTS	16 True positive	2 False positive	89% Precision	
	Model NTS	5 False negative	7 True negative	58% NPV	
		76% Sensitivity	78% Specificity	77% **Accuracy**	82% **F1 score**

NTS Score and Task Performance Metrics

The second rater completed ratings for 14 docking phases. The average values between the raters were used for the analysis, and ratings by one rater was used for cases without a second rater. The ICC between the two raters was 0.35 for the NTS ratings and 0.45 for the number of incidents. The docking duration had a positive correlation with overall NTS score and this relationship trended in significance ($r^2 = 0.15$, $p = 0.07$). Additionally, a negative correlation between the NTS score and number of incidents may be present ($r^2 = 0.13$, $p = 0.09$).

4.3.5 Objective NTS Model with Behavioral and Task Performance Metrics

The 19 behavioral metrics were used to construct a mixed effects model to predict the two task performance metrics. Table 4.8 and Table 4.9 summarizes the prediction of docking duration and number of incidents, respectively. No significant fixed effects were found for the docking duration model; however, eight fixed effects were significant in the prediction of incidences. The behavioral metric model was significantly different from the null models to predict duration ($\chi^2(18) = 30.95$, $p = 0.03$) and incidences ($\chi^2(19) = 61.56$, $p < 0.01$).

Table 4.8. Summary of mixed effects model of behavioral metrics on docking duration

Feature	**Estimate**	**Std. Error**	***p* - value**
Intercept	4.604	13.586	0.752
Surgeon-Circulating Callout	0.091	0.994	0.932
Surgeon-Circulating Request	-0.149	0.613	0.819
Surgeon-Circulating Confirmation	1.854	3.052	0.576
Surgeon-Circulating Non-verbal	-0.145	1.598	0.932
Surgeon-Technician Non-verbal	-0.078	0.931	0.937
Surgeon-Anesthesiologist Callout	-2.380	12.887	0.862
Surgeon-Anesthesiologist Request	2.990	7.409	0.707
Surgeon-Anesthesiologist Question	-3.306	17.210	0.857
Circulating-Surgeon Closed-Loop	0.847	3.721	0.831
Circulating-Surgeon Question	-1.319	2.557	0.633
Technician-Surgeon Request	0.136	0.406	0.754
Anesthesiologist-Surgeon Confirmation	1.091	2.725	0.709
Speech Mean Duration	-0.020	0.100	0.850
Speech Mean Pitch	3.574	79.177	0.966
Assisting % Time Close	-3.309	18.514	0.867
Assisting % Time Near	6.505	7.987	0.461
Circulating % Time Near	-4.269	8.256	0.632
Circulating % Time Far	0.687	1.182	0.592
Circulating Overall Mean Distance	4.604	13.586	0.752

Table 4.9. Summary of mixed effects model of behavioral metrics on number of incidents

Feature	Estimate	Std. Error	*p* - value
Intercept	10.150	9.096	0.408
Surgeon-Circulating Callout	-0.295	0.462	0.565
Surgeon-Circulating Request	0.537	0.300	0.172
Surgeon-Circulating Confirmation	-5.035	1.194	0.016*
Surgeon-Circulating Non-verbal	2.969	0.581	0.012*
Surgeon-Technician Non-verbal	-0.229	0.343	0.545
Surgeon-Anesthesiologist Callout	-19.780	4.989	0.018*
Surgeon-Anesthesiologist Request	7.207	3.158	0.088
Surgeon-Anesthesiologist Question	18.450	6.993	0.058
Circulating-Surgeon Closed-Loop	-0.246	1.990	0.913
Circulating-Surgeon Question	3.860	1.144	0.032*
Technician-Surgeon Request	0.199	0.197	0.394
Anesthesiologist-Surgeon Confirmation	-3.868	1.187	0.034*
Speech Mean Duration	0.005	0.062	0.940
Speech Mean Pitch	-127.100	28.060	0.020*
Assisting % Time Close	29.960	6.805	0.016*
Assisting % Time Near	-11.570	2.870	0.025*
Circulating % Time Near	-2.116	2.991	0.528
Circulating % Time Far	0.455	0.488	0.405
Circulating Overall Mean Distance	10.150	9.096	0.408

* indicates statistically significant ($p < 0.05$)

4.4 Discussion

The innovative technology for robotic-assisted procedures has been designed for better dexterity and visualization for surgeons; however, with the increased system complexity in the OR, there are shifts in NTS demands (Catchpole et al., 2016; Raison et al., 2017). Standard NTS assessments are currently observer- and checklist-based tools that have been developed to score NTS. Limitations of these evaluations exist such as the time-intensive nature and potential rater biases; therefore, objective NTS measures are needed. Behavioral metrics can be used as objective measures of NTS, and quantified task performance metrics can infer NTS of a clinical team and provide triangulation for the validity of the behavioral NTS metrics. In this study, these objective measures were utilized to construct models to predict NTS scores of surgeons using a standard evaluation tool. Specifically, a behavioral model composed of communication, speech, and proximity features was used to predict NTS score. This model was then used to estimate task performance metrics, thus building a fully objective NTS model.

4.4.1 Observational NTS Scores

Overall NTS scores – except for two observations – ranged between the acceptable and good range among the four surgeon participants. Interestingly, the lowest mean NTS was observed from Surgeon 1, who reported the most experience. Previous work has found significant decreases in decision-making and communication/teamwork construct scores as the number of years since fellowship increases (Gostlow et al., 2017). The authors hypothesized that this may be attributed to the decreases in explicit articulation of decision-making points and not seeking the opinions of team members during the procedure. Observations of Surgeon 1 agree with these inferences: there were fewer behaviors that were observed that were used to rate NTS and this participant rarely encouraged input from all team members. This alludes to factors that may explain the distribution of overall NTS scores.

Standard NTS assessments rely on raters observing explicit behaviors to be used for the observational assessment tools. Though this study did not focus on team familiarity, the experience of the surgical team, and especially of the assistant, may have influenced these ratings. For all Surgeon 1 cases, a senior resident assisted in the procedure; however, assistants during the other surgeons' procedures ranged from medical students to fellows (i.e., trainees that have completed residency and additional specialty training). As team familiarity has been found to influence NTS of surgical team members (Gillespie et al., 2017; Kang et al., 2015), the teamwork between Surgeon 1 and the assistant may have affected the surgeon's need to explicitly state specific communications for which the raters can observe and rate. In addition, differences in task and teamwork demands of a surgical phase may have influenced variations of NTS scores among the phases. For example, the docking phase typically requires the coordination of all team members: the surgeon guiding the positioning of the robot; the circulating nurse driving the robotic arms; and the assistant, technician, and anesthesiologist ensuring that the sterile field is clear for the robot placement. During this phase, the increase number of communications and behaviors increased events or actions for raters to assess NTS. On the contrary, surgeons often had less communication, thus less opportunities to extract NTS, during the critical phase as they were focusing on technical aspects of the operation.

Furthermore, this limited number of observed behaviors attributes to the small distribution of NTS scores within the data. No poor NTS behaviors (NOTSS score = 1) were noted by the raters in this study, and marginal behaviors (NOTSS score = 2) were seldom; a rater assigns these

scores for behaviors or actions that need remediation or considerable improvement (S. Yule et al., 2008). Though behaviors that may have influenced poor NTS may have occurred, implicit behaviors that could not be observed or inferred by raters may have been missed by the raters. Additionally, poor NTS may have been influenced by the increases in task demands and workload. Though not statistically significant, positive correlations between self-perceived effort, frustration, and situational stress with NTS were observed. This relationship may have been due to the Hawthorne effect: the surgeons' awareness of NTS observers may cause individuals to put more effort into the overall procedure and be cognizant of the presence of additional personnel in the OR (McCarney et al., 2007). The overall increase in demands could have contributed to their assessment of frustration during the procedure and may represent increases of surgeons' NTS behaviors. Concurrently, the negative relationship of self-perceived task complexity and NTS score may represent the prioritization of technical skills during a complex procedure. Surgeons focusing on the technical tasks often had decreases in communication, and with fewer opportunities to infer NTS, raters may have smaller data points to use when assigning an NTS score.

Though NTS construct scores were assigned and is considered as ordinal values, this work used overall NTS score as a continuous metric. The overall NTS score was calculated from the mean of the four NOTSS constructs, which was calculated as a mean of the 3 elements in each respective construct. By deriving overall NTS score as an average of 12 elements, the NTS score was an interval and parametric tests were applied (Norman, 2010). Moreover, for the docking phases that were rated by a second observer, there was minimal interrater agreement for NTS score and incidents. Though training cases were discussed to find calibrate the two raters, the low sample size may influence the low agreement. This thus emphasizes the need for objective NTS measurement to minimize variability.

4.4.2 Behavior Metrics and Models

Behavior metrics identified from the scoping review presented in Chapter 3 were utilized in addition to speech and proximity metrics. From the individual linear correlation, it was found that communication with team members not in the sterile field influence NTS. Specifically, 12 out of the 19 behavioral features included those with communications with the anesthesiologist or circulating nurse. During RAS procedures, the anesthesiologist is typically isolated, where there

is a physical barrier (i.e., drape) to section the sterile field and the individual is surrounded by differences monitors and the anesthesia station. Depending on the OR layout and robot position for the procedure, the anesthesiologist may not be in the field of view of the surgeon; they are often on opposite sides of the room when the surgeon is on the console. Due to this increase in spatial distance between the surgeon and anesthesiologist, interactions allowing the anesthesiologist to be aware of the operation flow are often noted in NTS ratings. In addition, the moderate correlation of NTS score with the %time the circulating nurse was far away from the surgeon suggests that there were less opportunities for observed for interactions to be used by raters for NTS scoring. Non-verbal communications with the surgeon to the circulating nurse and technician were found to be significantly correlated with NTS score. These behaviors, typically composed of gestures or movements, have been indicative of anticipation, which has been identified as a key component for team effectiveness (Annett et al., 2000; Sexton et al., 2018; B. Zheng et al., 2007). With surgeon interactions with the technician and circulating nurse in particular, non-verbal communication may increase due to non-verbal passing of instruments and exchanges or requests through gestures.

At least two features from each of the three metrics were statistically correlated with NTS score, which suggests that the joint application of these metrics can help predict NTS. From the two mixed effects models predicting either NTS score or incidences, the fixed effects of mean pitch and percent time the assistant was near the surgeon were large. This suggests that every 0.005 Hz increase of mean pitch is related to one NTS score and incident increase. An increase in 0.35% more time the assistant is near the surgeon is positively related to a NTS score increase; however, there is an inverse relationship that in every 11.6% decrease of time the assistant and surgeon are not in the near proximity, the incident count increases. As the assistant is in the near proximity to the surgeon, the surgeon may be completing positive behaviors (e.g., teaching) that may be inferred as positive NTS; however, when the surgeon is not within this range (e.g., when the surgeon is at the robot console and assistant is at the patient bed), incidents are more likely to occur. Furthermore, the significance of pitch provide insight into the surgeon's management of stressful stimuli. It has been found that a speaker's pitch changes as a response to stimulated cognitive demands (Lively et al., 1993; Mendoza & Carballo, 1998). The increase of incidents with pitch may represent the surgeon's internal management of increased environmental demands (e.g., they perceive that an incident may happen); thus, they may actively communicate more with team members, which are then observed by raters to improve NTS scores.

Non-linear behavior metric models are potentially better predictors of NTS. Due to the small distribution and sample size, classification models were built to predict exemplar behavior (NTS score = 4) and non-exemplar behavior (NTS Score <4). The three non-linear models have higher AUC than the linear model; although it is important to note the large confidence interval for all classifiers. However, the accuracy and F1 score of the best RF model suggests that non-linear models with NTS can better predict exemplar behavior than linear models. This suggests the need for considering non-linear relationships of the behavioral metrics with NTS, and implementation of non-linear models to predict NTS scores, by expanding from just exemplar versus non-exemplar behavior.

4.4.3 Task Performance Metrics and Models

Task performance metrics of time and incidents may be sensitive to changes in NTS scores. Though the linear model was not significant, the positive correlation between docking duration and NTS suggests that when surgeons spend longer times for docking the robot, there may be additional communication between the team that positively influences NTS score. Additionally, this non-significant relationship may represent that although increases in task completion time are associated with poor technical skills, this may not be the case for NTS. When a surgeon is spending additional time to teach a trainee, though this may increase overall operation time, this may be indicative of good NTS. Yet, it should be recalled that NTS should complement technical skill and efficiency, so a tradeoff between time and increased behaviors (e.g., too much time teaching or guiding trainees) should be explored. Furthermore, as expected, there was a negative relationship between the number of incidents and NTS score. This is likely since when NTS raters perform an evaluation, they consider the surgeon's reactions to an incident. Positive actions by the surgeon during unavoidable events are typically rated with higher NTS; however, poor responses to incidents that could have been avoidable or unavoidable are rated lower (Siu et al., 2016). Though the number of incidents involving all surgical team members were quantified, the subset of those that involved the surgeon were considered during NTS rating. Although non-significant associations between the task performance metrics and subjective NTS were found, trends show the potential of considering these metrics into predictive NTS models.

A fully objective model was developed for predicting task performance metrics with behavioral features. The behavioral metrics model is indicative of docking time and incidents;

however, overfitting of the model is likely. Since these models were developed for predicting task performance during the docking phase, 19 fixed effects were used to predict 34 observations. Yet, the strong correlations suggest that a fully objective NTS model can be achieved.

4.4.4 Limitations & Future Work

Several limitations existed in this work and addressing their implications will strengthen the development of objective measurement of NTS in surgery. First, the number of cases and participant surgeons should increase. Cases included in this study represented 4 male, right-hand-dominant surgeons within two specialties. These surgeons, in particular, did not display poor NTS behaviors. The consistent demographics of the participants may limit the generalizability of the models to all surgeons. Also due to the small sample size, it cannot be concluded that demographic influences (e.g., years of experiences) NTS score from this data. Although Surgeon 1's NTS score was significantly lower than the others, this may be due to individual differences of NTS and not solely based on experience. Furthermore, the performance of the NTS score prediction models are unknown from this work due to no ratings of poor NTS behaviors (NOTSS score = 1). Due to this limitation, the classification model was developed to predict exemplar or non-exemplar behavior. For these models, the unbalanced class distribution of the dataset may not have accurately identified the minority exemplar score, and resampling techniques can be implemented to improve the accuracy. Thus, to address the above limitations, an increase in sample size to include range of surgeons with varying skills and expertise is needed for increased generalizability and validity.

Additional confounding factors of NTS should also be explored and measured; specifically, team composition and familiarity. This study was completed at a large academic institution with similar team compositions. Trainees often participated in the operation as assisting surgeons and scheduling of OR staff was typically consistent. Roles of individuals and the surgeons' interactions with them can vary in different specialties and institutions, thus this composition should be considered when generalizing the data. Moreover, previous work has investigated the relationship of team familiarity and its influence on NTS, operative flow, and efficiency (Cohen et al., 2016; Mazzocco et al., 2009; Sexton et al., 2018; Weigl et al., 2018). Considering the expertise level of trainees as well as overall familiarity level of the team it warranted: this can provide insight in not only time to complete actions but shifts in communication modes from verbal to non-verbal to increase anticipation among the team members.

Finally, although frequency-based measures were used to develop the objective behavioral models, quantifications of these metrics still relied on raters which does not eliminate subjectivity. This was especially observed by the low ICC among the two raters for the NTS scores and incidences. The raters' respective biases influenced the variations among the identification and consideration of events that influenced NTS. For example, during the training session, one rater noted that the leadership construct was lowered by one particular action of the surgeon. This halo error, where the rating is based on one observation, introduces a systematic bias that may reduce accuracy (Feldman et al., 2012).

Future work on the sensor metrics are needed to gain validity evidence and expand their applications. In particular, this was the first time that the proximity sensors were deployed by the study team. Although validation tests in a laboratory setting were completed, additional work in verifying the accuracy of the results are needed *in situ*. For instance, during its initial implementation in the OR, data from the sensors were dropped due to the presence of the OR equipment as well as the continuous movement of the surgical team. This led to the missing data of the proximity features, which were then imputed for analysis in this work. In order to capture true behavior in the OR, the sensors' reliability and accuracy should be further tested. Additionally, the use of sensor-based metrics to capture both behavioral and physiological signals can address the challenge of summarizing NTS during hours-long surgery to one score and has the potential for real-time analysis and feedback. This use of sensors-based metrics can be applied for objective NTS measurement and drive the safety and effectiveness for patient care in the OR.

5. GENERAL DISCUSSION

This research investigated objective measures of non-technical skills (NTS) of surgeons using behavioral and physiological metrics in an operating room (OR). As research has grown on measuring the impact of NTS on surgical performance and patient care, there has been increased attention in the need to assess NTS for effective surgical practice (Agha et al., 2015). This dissertation aimed to answer two research questions on first understanding the current state of the literature on objective – or quantified – behavioral or physiological metrics of NTS for surgeons and then investigating if these measures can be applied in the OR to measure surgeons' NTS objectively.

The scoping review of the literature mapped the intersection between NTS, objective measures, the surgical environment, and surgeons. Findings from this work showed that literature of objective NTS measurements in the OR is fractured and there are further needs for the advancement of this field. Ten objective metrics were identified to be associated with NTS constructs, thus identifying the metrics to answer the first research question of this dissertation. Of these 10 summarized metrics, eight were frequency-based that required rater observation and analysis. Although these measures were quantified in the studies included in the review, rater biases and subjectivity exist; therefore, there is a need for additional evidence on the reliability of these frequency-based metrics. Furthermore, the use of signals from brain and eye activity were identified to be potential real-time, continuous measures. These physiological responses were measured in simulation environments and the potential for its application into the OR need to be investigated. However, implementation of their use in the OR is feasible, as technologies measuring such signals have been previously worn by surgical team members in the OR to measure cognitive workload or attention (Guru, Shafiei, et al., 2015; Koh et al., 2011). Finally, a critical gap was identified in that no studies were found to focus specifically on measuring and relating NTS with objective metrics. Although the resulting articles in the scoping review reported associations between NTS constructs and objective measures, no study investigated the direct connection between the two. Thus, the second research question proposed in this dissertation was critical to understand this relationship.

The second study presented in this dissertation addressed the gap that although objective NTS measurements were found in the current literature, no study predicted NTS with these metrics

for surgeons. Behavior metrics identified from the scoping review presented in Chapter 3 were utilized in addition to speech and proximity metrics for the prediction models. The quantified behavioral metrics were able to predict NTS score of surgeons, without the use of physiological metrics. This supports the understanding that current observational assessment techniques are largely communication based, and evaluation is centralized through observable behaviors that represent the interpersonal or infer the cognitive skills. Directional relationships between the behavioral metrics and overall NTS scores were obtained; however, further work is needed to build upon this understanding of the metrics with specific NTS behaviors. The implementation of measuring physiological signals and integrating the metrics into the objective NTS models may help improve the prediction of standard NTS scores. It should be also noted that although physiological metrics has been annotated by observers (e.g., eye gaze to infer situation awareness) in studies included in the scoping review, there is still a dependence on raters. Applications of technologies to measure physiological responses (e.g., brain activity and eye-tracking sensors) can allow for the removal of perceived metrics through self- or expert-raters, which can change the paradigm of measuring NTS. Thus, through this study, the second research question of this dissertation was answered: it was found that clinicians' NTS has the potential to be measured objectively in the clinical environment.

5.1.1 Limitations of Studies

As discussed in the respective chapters, limitations of the current studies in investing objective NTS assessments exist. Specifically, from the scoping review, it was found that no physiological measures have been applied in the OR to associate with NTS and the leadership construct was not correlated with objective metrics in the identified literature. Leadership of team members are critical to understand and measure, as leadership skills of individuals may affect team performance and patient safety in surgery (Patel et al., 2010). Moreover, factors that limited the generalizability of the findings from the objective models developed from data collected in the OR were two-fold: confounders such as team composition was not controlled and there was a limited distribution of subjective NTS scores among the participants. These limitations allude to the dynamic environment and system of a surgical operation, and additional investigations and studies to obtain construct validity are needed.

5.1.2 Guidelines for Addressing Limitations of Studies

Although this work found relationships between objective and observational surgeon NTS in the OR, further expansion of the findings from Chapter 3 and 4 are needed. Specifically, relationships of the behavioral and physiological metrics should be quantified in a controlled, simulated setting. A randomized control trial can be completed to obtain validity evidence; in particular, data supporting that behavioral and physiological objective metrics represent NTS can be obtained. With the ability of integrating reference-standard assessments (e.g., indices of skills that separate NTS), the completion of an experimental study that addresses limitations from work presented in this dissertation can verify that the objective metrics are measuring NTS behavior. The following section will provide a study design for this expansion study.

5.1.3 Study Design for Experimental Study

The purpose of this study design is to describe a simulation study to build validity evidence of objectively measuring NTS in clinical environments and understand the relationships between the objective metrics and NTS behavioral characteristics. This study will measure NTS through both objective and observational measures alongside reference-standard assessments that quantify NTS behavior. Specifically, this plan is designed to simulate acute care scenarios with integrated events that aims to elicit behaviors that demonstrate NTS from a participant (Cha et al., 2019; Peng et al., 2019). Assessment of NTS will address the following research hypotheses:

1. NTS of an individual can be measured objectively using physiological and/or behavioral metrics (e.g., cardiovascular, brain or eye activity, speech, or communication metrics)
2. NTS can be distinguished by scenarios that elicit poor/good NTS responses.

Study Design

An overview of the study design is shown in Figure 5.1. Participants will be randomized into either the control or intervention group, where the intervention is NTS training. From a systematic review of NTS training tools, it was reported that NTS trainings have positive effects, thus can guide the improvement of NTS of individuals and teams (Wood et al., 2017). For this reason, NTS training will be implemented in this study to increase the distribution of observational NTS scores. Prior to the completion of the first scenario, participants will complete a robust

assessment battery (e.g., established questionnaires of confounders such as leadership questionnaire) to capture possible individual differences such as leadership questionnaires (Avolio & Bass, 2004). After the completion of the first scenario, the intervention group will receive NTS training. The training will be composed of a didactic 30-minute workshop by a trained NTS rater focusing on increase awareness of NTS constructs and examples of good and poor behaviors (Pena et al., 2015; Riley et al., 2011). The control group will complete an alternative training focused on technical skills (e.g., completing Fundamentals of Laparoscopic Surgery Skills tasks) as an active control. After the trainings, a second scenario will be completed by participants, followed by post-scenario questionnaires and an NTS training questionnaire by the intervention group. The two scenarios completed by the participants will be counterbalanced for the scenarios.

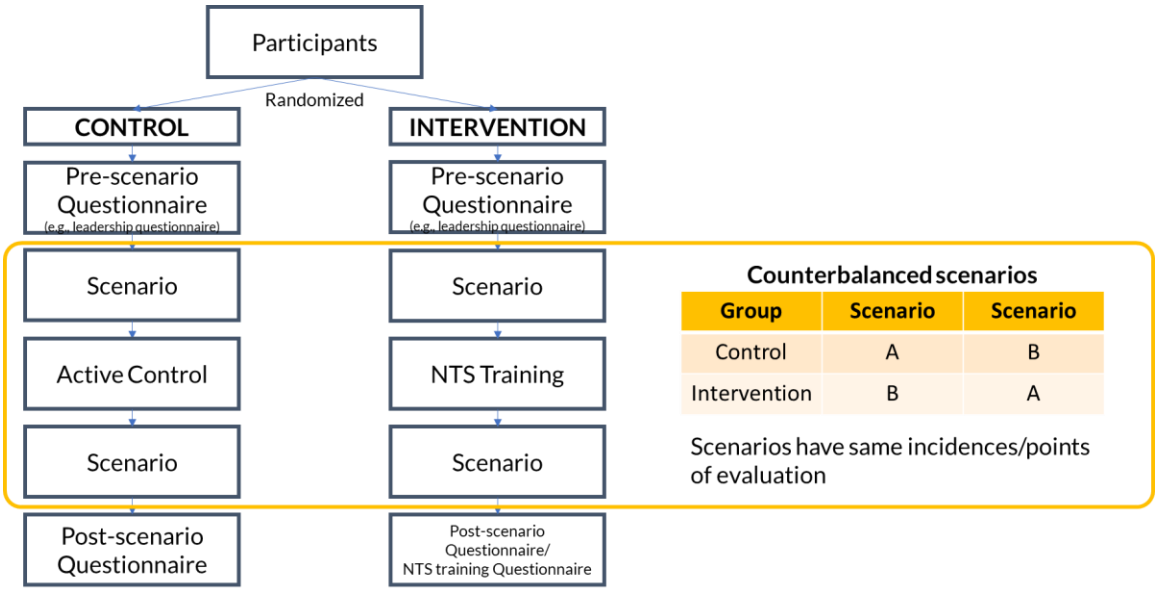

Figure 5.1. Study design of simulation study

The following section describes the requirements needed for the two 10-minute scenarios (Figure 5.2). Participants will have the role of the primary care provider responsible for determining a differential diagnosis of a simulated patient for different scenarios. A nurse confederate will also be embedded to facilitate the scenarios. Three incidences/events will be integrated into each scenario to measure the presence of an appropriate response by the participant: an introduction by the patient, need for a decision point, and an unexpected event. For each of these points of evaluation, assessors will note if the reference-standard behavior was observed.

First, the participant will be assessed on if an introduction was made to the patient at the start of the scenario prior to the patient making an introduction (i.e., yes or no an introduction was made). Next, it will be noted if the participant updates and informs the team of any progress or decisions after the decision point. For example, assessing if the participant informed both the patient and nurse on changes of a treatment plan. Finally, if the participant reacts to an unexpected event (i.e., alarm or phone ringing) will be measured. The completion of specific behaviors will be used as reference-standard metrics.

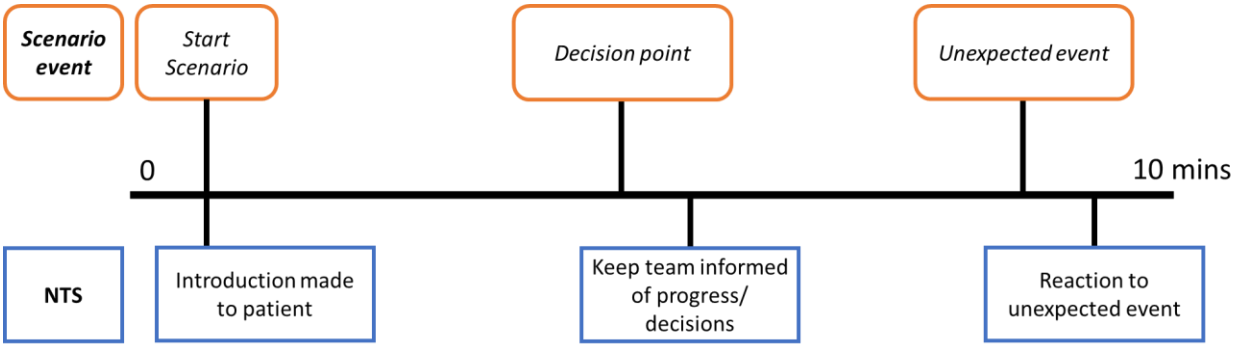

Figure 5.2. Scenario timeline with embedded events and NTS reference-standard behaviors

Metrics

Observational, objective, and reference-standard metrics will be measured. The standard NTS score will be obtained from assessment using the NOTSS tool, and reference-standard metrics, as described in the above section will be measured. Behavioral and physiological objective measures will be obtained from this scenario. As completed in the objective NTS measurement during RAS, communication, speech, and proximity metrics will be obtained. In addition, sensors to measure physiological signals will be deployed, specifically a mobile eye-tracker and heart rate sensor. Metrics focusing on conscious eye movements (e.g., number and duration of fixations, saccades, and areas of interest will be obtained) and involuntary pupillary response (e.g., blink rate) will be obtained. In addition, heart rate variability metrics such as the LF/HF ratio will be calculated to infer cognitive skills. It is noted that although brain activity metrics has been associated with NTS by the framework developed from the scoping review, implementing such sensors for scenarios that require continuous movement may not be feasible. Building on the

previous study, mixed effects models with the objective behavioral and physiological metrics will be built to predict observational NTS score and to classify if an appropriate response was observed for each event for reference-standard NTS.

Expected Results

It is expected that objective NTS models will be able to predict subjective NTS scores, as seen in the previous work shown in Chapter 4 of this dissertation. It is anticipated that the model will have better fit for predicting NTS score, and accuracy for classification models for predicting the presence of an appropriate behavior will be high. This is especially due to the addition of physiological metrics; however, this will increase the need for more sophisticated feature reduction techniques to best choose the features that model NTS. Additional analyses that can be completed in this study are comparisons of an objective model with just behavioral or just physiological metrics. A Likelihood Ratio Test can be completed with the null, behavioral, and physiological models. The objective model that is more significantly different than the null model will be determined as better predicting the reference-standard metric for NTS. Moreover, in this controlled simulation environment, directional relationships of NTS characteristics and the behavioral and physiological metrics can be determined.

5.1.4 Implications of Work and Expansion of Measurement

Immediate next steps of objective NTS measurement include expansion of measurement of NTS constructs and metrics as well as to measure different clinical teams and environments. The NTS construct of coping with stress/workload management has been often overlooked in standard NTS assessment tools due to the inability for raters to infer and assess cognitive skills. The use of objective physiological metrics may address this limitation of the rater-based tool to measure cognitive skills that were previously not evaluated. In parallel, additional metrics such as the pauses in communication can be quantified to gain insights on hesitations related to cognitive skills such as decision-making (Boomer & Dittmann, 1962; Maclay & Osgood, 1959).

This work focused on measuring surgeon NTS; however, each member of the surgical team contributes to patient care. The use of objective sensor-based measures has the potential to assess NTS of the surgical team concurrently in real-time. Near real-time applications can be used to

further investigate interventions to mitigate chances of errors and adverse events. Interventions executed by the team or the hospital system can be designed if NTS of individuals or teams are identified throughout surgery. For example, if it was found that the circulating nurse was out of the OR for large periods of time, this may be identified by the behavioral, proximity metrics. Due to this behavior's possible effect on the surgeon's and team's NTS, a notification could be given to the hospital staff to intervene to prevent poor NTS and possible detrimental events. Furthermore, this monitoring and intervention system can be applied not only for the OR, but for other high-stakes clinical environments such as the emergency room. For example, NTS has been previously identified to be significantly correlated with task performance in simulated trauma resuscitation (Briggs et al., 2015). Interestingly, this study reported NTS decline throughout the progression of a trauma care scenario, hypothesizing that required deviations due to unexpected events affects the procedural assessments that are needed for trauma care. These findings allude to the potential of NTS monitoring to clinical environments composed of procedural actions needed to complete a task and that are time sensitive.

Additionally, implementation of objective measures can be applied for evaluations of NTS training in medical education to minimize the need for time-intensive ratings by experts. As proposed in the simulation study, an NTS training program can be given to learners to gain understanding and improve their NTS. The expansion of assessments incorporating objective NTS can allow educators to have insights on cognitive skills that were previously not observable. Identification of specific, physiological responses can be used to guide and personalize teaching for trainees (e.g., teach optimal gaze strategies for situation awareness) to accelerate learning. Furthermore, application of these measures can allow for the monitoring of practicing surgeons and clinicians for continuous improvement. Generalized models and algorithms predicting behaviors related to poor NTS behaviors may be developed and deployed to enhance all practicing surgeons' and surgical teams' performance for better patient care and safety.

6. CONCLUSIONS

This research investigated objective measures of non-technical skills (NTS) of surgeons using behavioral and physiological metrics in an operating room (OR). There has been increased attention in the need to assess NTS, as they are critical alongside technical skills, for patient safety especially during surgery. Specifically, NTS measurement advances the understanding of the interpersonal and cognitive skills that are necessary in clinical environments for safety and effective performance. Limitations of current NTS evaluation tools include the need for trained raters and time-demand for training and assessment. This work was two-fold: a scoping review of the literature consolidating current objective NTS metrics of surgeons and developing models for this measurement in surgery.

A scoping review of the literature was completed to map the current state of applying the objective metrics to quantify surgeons' NTS. From the results, it was found that communication-based metrics were most used to quantify NTS. With this knowledge, measurement of behavioral metrics was deployed in the OR during robotic-assisted procedures to measure surgeon NTS. It was found that behavioral measures composed of communication, speech, and proximity metrics predicted subjective NTS. In addition, task performance metrics such as time and number of incidents during a procedure has the potential to be associated to the subjective NTS scores. Finally, it was found that the behavioral metrics can predict the task performance, showing the potential for a fully objective NTS measurement. Behavioral metrics has the potential to also overcome limitations of post-hoc task performance metrics, as they can be implemented for real-time NTS evaluation. Guidelines to address limitations of the current work was proposed, and future work include expansion of measurement metrics and applications of assessment to NTS training and the entire hospital system.

7. REFERENCES

Agha, R. A., Fowler, A. J., & Sevdalis, N. (2015). The role of non-technical skills in surgery. *Annals of Medicine and Surgery*, *4*(4), 422–427. https://doi.org/10.1016/j.amsu.2015.10.006

Ahmad, N., Hussein, A. A., Cavuoto, L., Sharif, M., Allers, J. C., Hinata, N., Ahmad, B., Kozlowski, J. D., Hashmi, Z., Bisantz, A., & Guru, K. A. (2016). Ambulatory movements, team dynamics and interactions during robot-assisted surgery. *BJU International*, *118*(1), 132–139. https://doi.org/10.1111/bju.13426

Ahmidi, N., Gao, Y., Béjar, B., Vedula, S. S., Khudanpur, S., Vidal, R., & Hager, G. D. (2013). String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, & N. Navab (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (pp. 26–33). Springer. https://doi.org/10.1007/978-3-642-40811-3_4

Al-Moteri, M. O., Symmons, M., Plummer, V., & Cooper, S. (2017). Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Computers in Human Behavior*, *66*, 52–66. https://doi.org/10.1016/j.chb.2016.09.022

Annett, J., Cunningham, D., & Mathias-Jones, P. (2000). A method for measuring team skills. *Ergonomics*, *43*(8), 1076–1094. https://doi.org/10.1080/00140130050084888

Antoniadis, S., Passauer-Baierl, S., Baschnegger, H., & Weigl, M. (2014). Identification and interference of intraoperative distractions and interruptions in operating rooms. *Journal of Surgical Research*, *188*(1), 21–29. https://doi.org/10.1016/j.jss.2013.12.002

Arora, S., Sevdalis, N., Nestel, D., Woloshynowych, M., Darzi, A., & Kneebone, R. (2010). The impact of stress on surgical performance: A systematic review of the literature. *Surgery*, *147*(3), 318-330.e6. https://doi.org/10.1016/j.surg.2009.10.007

Avolio, B. J., & Bass, B. M. (2004). Multifactor leadership questionnaire. *CA: Mind Garden*.

Baker, D. P., Amodeo, A. M., Krokos, K. J., Slonim, A., & Herrera, H. (2010). Assessing teamwork attitudes in healthcare: Development of the TeamSTEPPS teamwork attitudes questionnaire. *Qual Saf Health Care*, *19*(6), e49–e49. https://doi.org/10.1136/qshc.2009.036129

Baker, D. P., Gustafson, S., Beaubien, J., Salas, E., & Barach, P. (2005). Medical teamwork and patient safety: The evidence-based relation. *AHRQ Publication*, *5*(53), 1–64.

Baldwin, P. J., Paisley, A. M., & Brown, S. P. (1999). Consultant surgeons' opinion of the skills required of basic surgical trainees: Skills required of basic surgical trainees. *British Journal of Surgery*, *86*(8), 1078–1082. https://doi.org/10.1046/j.1365-2168.1999.01169.x

Barling, J., Akers, A., & Beiko, D. (2018). The impact of positive and negative intraoperative surgeons' leadership behaviors on surgical team performance. *The American Journal of Surgery*, *215*(1), 14–18. https://doi.org/10.1016/j.amjsurg.2017.07.006

Barzallo Salazar, M. J., Minkoff, H., Bayya, J., Gillett, B., Onoriode, H., Weedon, J., Altshuler, L., & Fisher, N. (2014). Influence of Surgeon Behavior on Trainee Willingness to Speak Up: A Randomized Controlled Trial. *Journal of the American College of Surgeons*, *219*(5), 1001–1007. https://doi.org/10.1016/j.jamcollsurg.2014.07.933

Bass, B. M. (1997). Does the transactional–transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist*, *52*(2), 130.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0*.

Bayramzadeh, S., Joseph, A., San, D., Khoshkenar, A., Taaffe, K., Jafarifiroozabadi, R., Neyens, D. M., & Group, R. O. S. (2018). The impact of operating room layout on circulating nurse's work patterns and flow disruptions: A behavioral mapping study. *HERD: Health Environments Research & Design Journal*, *11*(3), 124–138.

Bedny, G., & Meister, D. (1999). Theory of activity and situation awareness. *International Journal of Cognitive Ergonomics*, *3*(1), 63–72.

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., Zivkovic, V. T., Popovic, M. V., & Olmstead, R. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction*, *17*(2), 151–170.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). *EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks*. *78*(5), 14.

Bezemer, J., Murtagh, G., Cope, A., & Kneebone, R. (2016). Surgical decision making in a teaching hospital: A linguistic analysis: Surgical decision making. *ANZ Journal of Surgery*, *86*(10), 751–755. https://doi.org/10.1111/ans.12824

Biebuyck, J. F., Weinger, M. B., & Englund, C. E. (1990). Ergonomic and human factors affecting anesthetic vigilance and monitoring performance in the operating room environment. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *73*(5), 995–1021.

Boet, S., Larrigan, S., Martin, L., Liu, H., Sullivan, K. J., & Etherington, N. (2018). Measuring non-technical skills of anaesthesiologists in the operating room: A systematic review of assessment tools and their measurement properties. *British Journal of Anaesthesia*, *121*(6), 1218–1226. https://doi.org/10.1016/j.bja.2018.07.028

Boomer, D. S., & Dittmann, A. T. (1962). Hesitation pauses and juncture pauses in speech. *Language and Speech*, *5*(4), 215–220.

Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., Caltagirone, C., Kong, W., Wei, D., & Zhou, Z. (2012). Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices. *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 6442–6445.

Bousmalis, K., Mehu, M., & Pantic, M. (2009). Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–9.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing Communication Sequences for Team Training Needs Assessment. *Human Factors*, *40*(4), 672–679. https://doi.org/10.1518/001872098779649265

Braddock, C., Hudak, P. L., Feldman, J. J., Bereknyei, S., Frankel, R. M., & Levinson, W. (2008). "Surgery Is Certainly One Good Option": Quality and Time-Efficiency of Informed Decision-Making in Surgery. *The Journal of Bone and Joint Surgery. American Volume.*, *90*(9), 1830–1838. https://doi.org/10.2106/JBJS.G.00840

Briggs, A., Raja, A. S., Joyce, M. F., Yule, S. J., Jiang, W., Lipsitz, S. R., & Havens, J. M. (2015). The role of nontechnical skills in simulated trauma resuscitation. *Journal of Surgical Education*, *72*(4), 732–739.

Brown, I. D. (2002). A review of the'looked but failed to see'accident causation factor. *Behavioural Research in Road Safety: Eleventh Seminar*.

Buuren, S. van, & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.

Camm, A. J., Malik, M., Bigger, J. T., Breithardt, G., Cerutti, S., Cohen, R. J., Coumel, P., Fallen, E. L., Kennedy, H. L., & Kleiger, R. E. (1996). *Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology*.

Carswell, C. M., Clarke, D., & Seales, W. B. (2005). Assessing Mental Workload During Laparoscopic Surgery. *Surgical Innovation*, *12*(1), 80–90. https://doi.org/10.1177/155335060501200112

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. *Oncology Nursing Forum*, *41*(5), 545–547.

Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M., & Pecchia, L. (2015). Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control*, *18*, 370–377.

Catchpole, K., Bisantz, A., Hallbeck, M. S., Weigl, M., Randell, R., Kossack, M., & Anger, J. T. (2019). Human factors in robotic assisted surgery: Lessons from studies 'in the Wild.' *Applied Ergonomics*, *78*, 270–276.

Catchpole, K., Giddings, A. E. B., Wilkinson, M., Hirst, G., Dale, T., & de Leval, M. R. (2007). Improving patient safety by identifying latent failures in successful operations. *Surgery*, *142*(1), 102–110. https://doi.org/10.1016/j.surg.2007.01.033

Catchpole, K., Perkins, C., Bresee, C., Solnik, M. J., Sherman, B., Fritch, J., Gross, B., Jagannathan, S., Hakami-Majd, N., Avenido, R., & Anger, J. T. (2016). Safety, efficiency and learning curves in robotic surgery: A human factors analysis. *Surgical Endoscopy*, *30*(9), 3749–3761. https://doi.org/10.1007/s00464-015-4671-2

Cha, J. S., Anton, N. E., Mizota, T., Hennings, J. M., Rendina, M. A., Stanton-Maxey, K., Ritter, H. E., Stefanidis, D., & Yu, D. (2019). Use of non-technical skills can predict medical student performance in acute care simulated scenarios. *The American Journal of Surgery*, *217*(2), 323–328.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028

Cheriyan, S., Mowery, H., Ruckle, D., Keheila, M., Myklak, K., Alysouf, M., Atiga, C., Khuri, J., Khater, N., Faaborg, D., Ruckle, H. C., Baldwin, D. D., & Baldwin, D. D. (2016). The Impact of Operating Room Noise Upon Communication During Percutaneous Nephrostolithotomy. *Journal of Endourology*, *30*(10), 1062–1066. https://doi.org/10.1089/end.2016.0498

Clark, H. H. (1996). *Using language*. Cambridge university press.

Cohen, T. N., Cabrera, J. S., Sisk, O. D., Welsh, K. L., Abernathy, J. H., Reeves, S. T., Wiegmann, D. A., Shappell, S. A., & Boquet, A. J. (2016). Identifying workflow disruptions in the cardiovascular operating room. *Anaesthesia*, *71*(8), 948–954.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. *Team Cognition: Understanding the Factors That Drive Process and Performance*, 83–106.

Cooper, J. B. (1984). Toward prevention of anesthetic mishaps. *International Anesthesiology Clinics*, *22*(2), 167–184.

Cooper, J. B., Newbower, R. S., & Kitz, R. J. (1984). An analysis of major errors and equipment failures in anesthesia management: Considerations for prevention and detection. *Anesthesiology*, *60*(1), 34–42.

Cooper, S., Endacott, R., & Cant, R. (2010). Measuring non-technical skills in medical emergency care: A review of assessment measures. *Open Access Emergency Medicine : OAEM*, *2*, 7–16.

Cooper, W. O., Spain, D. A., Guillamondegui, O., Kelz, R. R., Domenico, H. J., Hopkins, J., Sullivan, P., Moore, I. N., Pichert, J. W., Catron, T. F., Webb, L. E., Dmochowski, R. R., & Hickson, G. B. (2019). Association of Coworker Reports About Unprofessional Behavior by Surgeons With Surgical Complications in Their Patients. *JAMA Surgery*, *154*(9), 828. https://doi.org/10.1001/jamasurg.2019.1738

Cristancho, S. M., Vanstone, M., Lingard, L., LeBel, M.-E., & Ott, M. (2013). When surgeons face intraoperative challenges: A naturalistic model of surgical decision making. *The American Journal of Surgery*, *205*(2), 156–162.

Cunningham, S., Chellali, A., Banez, J., & Cao, C. G. L. (2012). Design of a Spatial Aid for Communication in Robotic Surgery. *Volume 2: Applied Fluid Mechanics; Electromechanical Systems and Mechatronics; Advanced Energy Systems; Thermal Engineering; Human Factors and Cognitive Engineering*, 847–854. https://doi.org/10.1115/ESDA2012-82804

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

de Winter, J. C., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, *21*(1), 99–111.

Denzin, N. K. (1978). Triangulation: A case for methodological evaluation and combination. *Sociological Methods*, 339–357.

Desvergez, A., Winer, A., Gouyon, J.-B., & Descoins, M. (2019). An observational study using eye tracking to assess resident and senior anesthetists' situation awareness and visual perception in postpartum hemorrhage high fidelity simulation. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0221515

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1838.

Dias, R. D., Ngo-Howard, M. C., Boskovski, M. T., Zenati, M. A., & Yule, S. (2018). Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *BJS*, *105*(5), 491–501. https://doi.org/10.1002/bjs.10795

Dias, R. D., Zenati, M. A., Stevens, R., Gabany, J. M., & Yule, S. (2019). Physiological synchronization and entropy as measures of team cognitive load. *Journal of Biomedical Informatics*, *96*, 103250. https://doi.org/10.1016/j.jbi.2019.103250

Dixon, M. L., & Christoff, K. (2014). The lateral prefrontal cortex and complex value-based learning and decision making. *Neuroscience & Biobehavioral Reviews*, *45*, 9–18.

Duchowski, A. T. (2007). Eye tracking methodology. *Theory and Practice*, *328*(614), 2–3.

Duncan, S., & Fiske, D. W. (2015). *Face-to-face interaction: Research, methods, and theory*. Routledge.

Echeverria, V., Martinez-Maldonado, R., Power, T., Hayes, C., & Shum, S. B. (2018). Where Is the Nurse? Towards Automatically Visualising Meaningful Team Movement in Healthcare Education. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 74–78). Springer International Publishing.

Endsley, M. R. (1988a). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*, 97–101.

Endsley, M. R. (1988b). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, 789–795.

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279–286. https://doi.org/10.1002/chp.21156

Fernandez, R., Shah, S., Rosenman, E. D., Kozlowski, S. W. J., Parker, S. H., & Grand, J. A. (2017). Developing team cognition: A role for simulation. *Simulation in Healthcare : Journal of the Society for Simulation in Healthcare*, *12*(2), 96–103. https://doi.org/10.1097/SIH.0000000000000200

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' Non-Technical Skills (ANTS): Evaluation of a behavioural marker system. *British Journal of Anaesthesia*, *90*(5), 580–588.

Fletcher, J. (2010). The Prosody of Speech: Timing and Rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 521–602). Blackwell Publishing Ltd. https://doi.org/10.1002/9781444317251.ch15

Flin, R., Martin, L., Goeters, K.-M., Hörmann, H.-J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. In D. Harris & H. C. Muir (Eds.), *Human Factors and Aerospace Safety* (1st ed., pp. 97–119). Routledge. https://doi.org/10.4324/9781315194035-1

Flin, R., & O'Connor, P. (2017). *Safety at the sharp end: A guide to non-technical skills*. CRC Press.

Flin, R., Youngson, G. G., & Yule, S. (2015). *Enhancing surgical performance: A primer in non-technical skills*. CRC press.

Flin, R., Youngson, G., & Yule, S. (2007). How do surgeons make intraoperative decisions? *Quality & Safety in Health Care*, *16*(3), 235–239. https://doi.org/10.1136/qshc.2006.020743

Frankel, A., Gardner, R., Maynard, L., & Kelly, A. (2007). Using the communication and teamwork skills (CATS) assessment to measure health care team performance. *The Joint Commission Journal on Quality and Patient Safety*, *33*(9), 549–558.

Gaba, D. M. (1989). Human error in anesthetic mishaps. *International Anesthesiology Clinics, 27*(3), 137–147.

Gawande, A. A., Zinner, M. J., Studdert, D. M., & Brennan, T. A. (2003). Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*, *133*(6), 614–621.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: A Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review*, *23*(4), 523–552. https://doi.org/10.1007/s10648-011-9174-7

Gillespie, B. M., Chaboyer, W., & Murray, P. (2010). Enhancing Communication in Surgery Through Team Training Interventions: A Systematic Literature Review. *AORN Journal*, *92*(6), 642–657. https://doi.org/10.1016/j.aorn.2010.02.015

Gillespie, B. M., Harbeck, E., Kang, E., Steel, C., Fairweather, N., & Chaboyer, W. (2017). Correlates of non-technical skills in surgery: A prospective study. *BMJ Open*, *7*(1), e014480. https://doi.org/10.1136/bmjopen-2016-014480

Gordon, M., Darbyshire, D., & Baker, P. (2012). Non-technical skills training to enhance patient safety: A systematic review. *Medical Education, 46*(11), 1042–1054. https://doi.org/10.1111/j.1365-2923.2012.04343.x

Gordon, M., Farnan, J., Grafton-Clarke, C., Ahmed, R., Gurbutt, D., McLachlan, J., & Daniel, M. (2019). Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide No. 54. *Medical Teacher*, *41*(7), 732–745. https://doi.org/10.1080/0142159X.2018.1562166

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. J., & Cooke, N. J. (2003). Evaluation of Latent Semantic Analysis-based measures of team communications content. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*, 424–428.

Gostlow, H., Marlow, N., Thomas, M. J. W., Hewett, P. J., Kiermeier, A., Babidge, W., Altree, M., Pena, G., & Maddern, G. (2017). Non-technical skills of surgical trainees and experienced surgeons. *British Journal of Surgery*, *104*(6).

Grawunder, S., & Winter, B. (2010). Acoustic correlates of politeness: Prosodic and voice quality measures in polite and informal speech of Korean and German speakers. *Speech Prosody 2010-Fifth International Conference*.

Grice, H. P., Cole, P., & Morgan, J. L. (1975). Logic and conversation. *1975*, 41–58.

Grundgeiger, T., Wurmb, T., & Happel, O. (2015). Eye Tracking in Anesthesiology: Literature Review, Methodological Issues, and Research Topics. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 493–497. https://doi.org/10.1177/1541931215591106

Guru, K. A., Esfahani, E. T., Raza, S. J., Bhat, R., Wang, K., Hammond, Y., Wilding, G., Peabody, J. O., & Chowriappa, A. J. (2015). Cognitive skills assessment during robot-assisted surgery: Separating the wheat from the chaff. *BJU International*, *115*(1), 166–174.

Guru, K. A., Shafiei, S. B., Khan, A., Hussein, A. A., Sharif, M., & Esfahani, E. T. (2015). Understanding Cognitive Performance During Robot-Assisted Surgery. *Urology*, *86*(4), 751–757. https://doi.org/10.1016/j.urology.2015.07.028

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 301–310.

Hansen, A. L., Johnsen, B. H., Sollers, J. J., Stenvik, K., & Thayer, J. F. (2004). Heart rate variability and its relation to prefrontal cognitive function: The effects of training and detraining. *European Journal of Applied Physiology*, *93*(3), 263–272.

Härgestam, M., Hultin, M., Brulin, C., & Jacobsson, M. (2016). Trauma team leaders' non-verbal communication: Video registration during trauma team training. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *24*(1), 37. https://doi.org/10.1186/s13049-016-0230-7

Härgestam, M., Lindkvist, M., Brulin, C., Jacobsson, M., & Hultin, M. (2013). Communication in interdisciplinary teams: Exploring closed-loop communication during in situ trauma team training. *BMJ Open*, *3*(10), e003525. https://doi.org/10.1136/bmjopen-2013-003525

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Model assessment and selection. In *The elements of statistical learning* (pp. 219–259). Springer.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512–521.

Hazlehurst, B., McMullen, C. K., & Gorman, P. N. (2007). Distributed cognition in the heart room: How situation awareness arises from coordinated communications during cardiac surgery. *Journal of Biomedical Informatics*, *40*(5), 539–551. https://doi.org/10.1016/j.jbi.2007.02.001

Healey, A. N. (2004). Developing observational measures of performance in surgical teams. *Quality and Safety in Health Care*, *13*(suppl_1), i33–i40. https://doi.org/10.1136/qshc.2004.009936

Helmreich, R. L., & Schaefer, H.-G. (1994). *Team performance in the operating room.*

Hersey, P., Blanchard, K. H., & Natemeyer, W. E. (1979). Situational leadership, perception, and the impact of power. *Group & Organization Studies*, *4*(4), 418–428.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*(4), 46–55.

Hjortdahl, M., Ringen, A. H., Naess, A.-C., & Wisborg, T. (2009). Leadership is the essential non-technical skill in the trauma team—Results of a qualitative study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *17*(1), 48. https://doi.org/10.1186/1757-7241-17-48

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., & O'Cathain, A. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291.

Horwitz, I. B., Horwitz, S. K., Daram, P., Brandt, M. L., Charles Brunicardi, F., & Awad, S. S. (2008). Transformational, Transactional, and Passive-Avoidant Leadership Characteristics of a Surgical Resident Cohort: Analysis Using the Multifactor Leadership Questionnaire and Implications for Improving Surgical Education Curriculums. *Journal of Surgical Research*, *148*(1), 49–59. https://doi.org/10.1016/j.jss.2008.03.007

Hu, Y.-Y., Parker, S. H., Lipsitz, S. R., Arriaga, A. F., Peyre, S. E., Corso, K. A., Roth, E. M., Yule, S., & Greenberg, C. C. (2016). Surgeons' Leadership Styles and Team Behavior in the Operating Room. *Journal of the American College of Surgeons*, *222*(1), 41–51. https://doi.org/10.1016/j.jamcollsurg.2015.09.013

Hull, L., Arora, S., Aggarwal, R., Darzi, A., Vincent, C., & Sevdalis, N. (2012). The Impact of Nontechnical Skills on Technical Performance in Surgery: A Systematic Review. *Journal of the American College of Surgeons*, *214*(2), 214–230. https://doi.org/10.1016/j.jamcollsurg.2011.10.016

Jacobs, J., Hwang, G., Curran, T., & Kahana, M. J. (2006). EEG oscillations and recognition memory: Theta correlates of memory retrieval and decision making. *NeuroImage*, *32*(2), 978–987. https://doi.org/10.1016/j.neuroimage.2006.02.018

Jung, J. J., Borkhoff, C. M., Jüni, P., & Grantcharov, T. P. (2018). Non-Technical Skills for Surgeons (NOTSS): Critical appraisal of its measurement properties. *The American Journal of Surgery*, *216*(5), 990–997. https://doi.org/10.1016/j.amjsurg.2018.02.021

Just, M. A., & Carpenter, P. A. (1975). *Eye fixations and cognitive processes.*

Kahneman, D. (2011). *Thinking, fast and slow* (Vol. 1). Farrar, Straus and Giroux New York.

Kang, E., Massey, D., & Gillespie, B. M. (2015). Factors that influence the non-technical skills performance of scrub nurses: A prospective study. *Journal of Advanced Nursing*, *71*(12), 2846–2857.

Kazi, S., Khaleghzadegan, S., Dinh, J. V., Shelhamer, M. J., Sapirstein, A., Goeddel, L. A., Chime, N. O., Salas, E., & Rosen, M. A. (2019). Team Physiological Dynamics: A Critical Review. *Human Factors*, 0018720819874160. https://doi.org/10.1177/0018720819874160

Keller, S., Tschan, F., Beldi, G., Kurmann, A., Candinas, D., & Semmer, N. K. (2016). Noise peaks influence communication in the operating room. An observational study. *Ergonomics*, *59*(12), 1541–1552. https://doi.org/10.1080/00140139.2016.1159736

Kesten, K. S. (2011). Role-play using SBAR technique to improve observed communication skills in senior nursing students. *Journal of Nursing Education*, *50*(2), 79–87.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., & Shope, S. M. (2001). Automating measurement of team cognition through analysis of communication data. *Usability Evaluation and Interface Design*, 1382–1386.

Kiekel, P. A., Gorman, J. C., & Cooke, N. J. (2017). Communication as team-level cognitive processing. In *Macrocognition in teams* (pp. 51–64). CRC Press.

Klein, G. A. (1993). *A recognition-primed decision (RPD) model of rapid decision making*. Ablex Publishing Corporation New York.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*, *29*(2–3), 169–195.

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., & Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neuroscience Letters*, *244*(2), 73–76.

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Koch, A., Burns, J., Catchpole, K., & Weigl, M. (2020). Associations of workflow disruptions in the operating room with surgical outcomes: A systematic review and narrative synthesis. *BMJ Quality & Safety*, bmjqs-2019-010639. https://doi.org/10.1136/bmjqs-2019-010639

Koh, R. Y., Park, T., Wickens, C. D., Ong, L. T., & Chia, S. N. (2011). Differences in attentional strategies by novice and experienced operating theatre scrub nurses. *Journal of Experimental Psychology: Applied, 17*(3), 233.

Kohn, L. T., Corrigan, J., & Donaldson, M. S. (2000). *To err is human: Building a safer health system* (Vol. 6). National academy press Washington, DC.

Korkiakangas, T., Weldon, S.-M., Bezemer, J., & Kneebone, R. (2014). Nurse–surgeon object transfer: Video analysis of communication and situation awareness in the operating theatre. *International Journal of Nursing Studies*, *51*(9), 1195–1206. https://doi.org/10.1016/j.ijnurstu.2014.01.007

Kriston, L., Scholl, I., Hölzel, L., Simon, D., Loh, A., & Härter, M. (2010). The 9-item Shared Decision Making Questionnaire (SDM-Q-9). Development and psychometric properties in a primary care sample. *Patient Education and Counseling*, *80*(1), 94–99. https://doi.org/10.1016/j.pec.2009.09.034

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345–360.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Laplante, D., & Ambady, N. (2003). On how things are said: Voice tone, voice intensity, verbal content, and perceptions of politeness. *Journal of Language and Social Psychology*, *22*(4), 434–441.

Leff, D. R., Elwell, C. E., Orihuela-Espina, F., Atallah, L., Delpy, D. T., Darzi, A. W., & Yang, G. Z. (2008). Changes in prefrontal cortical behaviour depend upon familiarity on a bimanual co-ordination task: An fNIRS study. *NeuroImage*, *39*(2), 805–813. https://doi.org/10.1016/j.neuroimage.2007.09.032

Leff, D. R., Yongue, G., Vlaev, I., Orihuela-Espina, F., James, D., Taylor, M. J., Athanasiou, T., Dolan, R., Yang, G.-Z., & Darzi, A. (2017). "Contemplating the Next Maneuver": Functional Neuroimaging Reveals Intraoperative Decision-making Strategies. *Annals of Surgery*, *265*(2), 320–330. https://doi.org/10.1097/SLA.0000000000001651

Lehto, M. R., & Nah, F. (2006). Decision-Making Models and Decision Support. *Handbook of Human Factors and Ergonomics*, 191–242.

Leuschner, S., Leuschner, M., Kropf, S., & Niederbichler, A. D. (2018). Non-technical skills training in the operating theatre: A meta-analysis of patient outcomes. *The Surgeon*. https://doi.org/10.1016/j.surge.2018.07.001

Lingard, L., Espin, S., Whyte, S., Regehr, G., Baker, G. R., Reznick, R., Bohnen, J., Orser, B., Doran, D., & Grober, E. (2004). Communication failures in the operating room: An observational classification of recurrent types and effects. *BMJ Quality & Safety*, *13*(5), 330–334.

Lingard, L., & Haber, R. J. (1999). Teaching and learning communication in medicine: A rhetorical approach. *Academic Medicine: Journal of the Association of American Medical Colleges*, *74*(5), 507–510.

Lingard, L., Reznick, R., Espin, S., Regehr, G., & DeVito, I. (2002). Team communications in the operating room: Talk patterns, sites of tension, and implications for novices. *Academic Medicine*, *77*(3), 232–237.

Lively, S. E., Pisoni, D. B., Van Summers, W., & Bernacki, R. H. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America*, *93*(5), 2962–2973.

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, *24*(1), 71–89.

Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology*, *113*, 83–90. https://doi.org/10.1016/j.biopsycho.2015.11.013

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835–1838.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19–44.

Madani, A., Vassiliou, M. C., Watanabe, Y., Al-Halabi, B., Al-Rowais, M. S., Deckelbaum, D. L., Fried, G. M., & Feldman, L. S. (2017). What Are the Principles That Guide Behaviors in the Operating Room?: Creating a Framework to Define and Measure Performance. *Annals of Surgery*, *265*(2), 255–267. https://doi.org/10.1097/SLA.0000000000001962

Makary, M. A., Sexton, J. B., Freischlag, J. A., Holzmueller, C. G., Millman, E. A., Rowen, L., & Pronovost, P. J. (2006). Operating Room Teamwork among Physicians and Nurses: Teamwork in the Eye of the Beholder. *Journal of the American College of Surgeons*, *202*(5), 746–752. https://doi.org/10.1016/j.jamcollsurg.2006.01.017

Malandro, L. A., & Barker, L. L. (1983). *Nonverbal communication*. Addison Wesley Publishing Company.

Malik, M., & Camm, A. J. (1990). Heart rate variability. *Clinical Cardiology*, *13*(8), 570–576.

Marshall, S. P. (2000). *Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity*.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*, 7–7.

Mayer, C. M., Cluff, L., Lin, W.-T., Willis, T. S., Stafford, R. E., Williams, C., Saunders, R., Short, K. A., Lenfestey, N., & Kane, H. L. (2011). Evaluating efforts to optimize TeamSTEPPS implementation in surgical and pediatric intensive care units. *The Joint Commission Journal on Quality and Patient Safety*, *37*(8), 365–AP3.

Mazzocco, K., Petitti, D. B., Fong, K. T., Bonacum, D., Brookey, J., Graham, S., Lasky, R. E., Sexton, J. B., & Thomas, E. J. (2009). Surgical team behaviors and patient outcomes. *The American Journal of Surgery*, *197*(5), 678–685. https://doi.org/10.1016/j.amjsurg.2008.03.002

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: A randomised, controlled trial. *BMC Medical Research Methodology*, *7*(1), 30.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

McMullan, R. D., Urwin, R., Sunderland, N., & Westbrook, J. (2020). Observational Tools That Quantify Nontechnical Skills in the Operating Room: A Systematic Review. *Journal of Surgical Research*, *247*, 306–322. https://doi.org/10.1016/j.jss.2019.10.012

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stressby means of cognitive workload tasks. *Journal of Voice*, *12*(3), 263–273.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., & Kramer, A. F. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cognitive Brain Research*, *12*(3), 467–473.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Miller, G. A., & Glucksberg, S. (1988). Psycholinguistic aspects of pragmatics and semantics. *Stevens' Handbook of Experimental Psychology*, *2*, 417–472.

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS System: Reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Quality and Safety in Health Care*, *18*(2), 104–108. https://doi.org/10.1136/qshc.2007.024760

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., & Youngson, G. (2011). Thinking ahead of the surgeon. An interview study to identify scrub nurses' non-technical skills. *International Journal of Nursing Studies*, *48*(7), 818–828. https://doi.org/10.1016/j.ijnurstu.2010.11.005

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., & Youngson, G. (2012). Evaluation of the scrub practitioners' list of intraoperative non-technical skills (SPLINTS) system. *International Journal of Nursing Studies*, *49*(2), 201–211.

Moore, A., Butt, D., Ellis-Clarke, J., & Cartmill, J. (2010). Linguistic analysis of verbal and non-verbal communication in the operating room. *ANZ Journal of Surgery*, *80*(12), 925–929.

Moorthy, K., Munz, Y., Adams, S., Pandey, V., & Darzi, A. (2006). Self-assessment of performance among surgical trainees during simulated procedures in a simulated operating theater. *The American Journal of Surgery*, *192*(1), 114–118. https://doi.org/10.1016/j.amjsurg.2005.09.017

Moorthy, K., Munz, Y., Sarker, S. K., & Darzi, A. (2003). Objective assessment of technical skills in surgery. *Bmj*, *327*(7422).

Moss, J., & Xiao, Y. (2004). Improving Operating Room Coordination: Communication Pattern Assessment. *JONA: The Journal of Nursing Administration*, *34*(2), 93–100. https://doi.org/10.1097/00005110-200402000-00008

Nemani, A., Kruger, U., Cooper, C. A., Schwaitzberg, S. D., Intes, X., & De, S. (2019). Objective assessment of surgical skill transfer using non-invasive brain imaging. *Surgical Endoscopy*, *33*(8), 2485–2494. https://doi.org/10.1007/s00464-018-6535-z

Nemani, A., Kruger, U., Intes, X., & De, S. (2017). Increased Sensitivity in Discriminating Surgical Motor Skills Using Prefrontal Cortex Activation over Established Metrics. *Optics in the Life Sciences Congress*, JTu4A.11. https://doi.org/10.1364/BODA.2017.JTu4A.11

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Nyssen, A.-S., & Blavier, A. (2010). Integrating Collective Work Aspects in the Design Process: An Analysis Case Study of the Robotic Surgery Using Communication as a Sign of Fundamental Change. In P. Palanque, J. Vanderdonckt, & M. Winckler (Eds.), *Human Error, Safety and Systems Development* (Vol. 5962, pp. 18–27). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11750-3_2

O'Meara, P., Munro, G., Williams, B., Cooper, S., Bogossian, F., Ross, L., Sparkes, L., Browning, M., & McClounan, M. (2015). Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper. *International Emergency Nursing*, *23*(2), 94–99. https://doi.org/10.1016/j.ienj.2014.11.001

Orasanu, J., & Fischer, U. (1997). Finding decisions in natural environments: The view from the cockpit. *Naturalistic Decision Making*, 343–357.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59. https://doi.org/10.1002/hbm.20131

Panesar, S. S., Carson-Stevens, A., Mann, B. S., Bhandari, M., & Madhok, R. (2012). Mortality as an indicator of patient safety in orthopaedics: Lessons from qualitative analysis of a database of medical errors. *BMC Musculoskeletal Disorders*, *13*(1), 93.

Parasuraman, R., & Wilson, G. F. (2008). Putting the brain to work: Neuroergonomics past, present, and future. *Human Factors*, *50*(3), 468–474.

Paris, C. R., Salas, E., & Cannon-Bowers, J. A. (2000). Teamwork in multi-person systems: A review and analysis. *Ergonomics*, *43*(8), 1052–1075.

Parker, S. H., Flin, R., McKinley, A., & Yule, S. (2013). The Surgeons' Leadership Inventory (SLI): A taxonomy and rating system for surgeons' intraoperative leadership skills. *The American Journal of Surgery*, *205*(6), 745–751.

Parker, S. H., Yule, S., Flin, R., & McKinley, A. (2012). Surgeons' leadership in the operating room: An observational study. *The American Journal of Surgery*, *204*(3), 347–354.

Parker-Raley, J., Mottet, T. P., Lawson, K. A., Duzinski, S. V., Cerroni, A., & Mercado, M. (2012). Investigating pediatric trauma team communication effectiveness phase one: The development of the assessment of pediatric resuscitation communication. *Journal of Communication in Healthcare*, *5*(2), 102–115.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1–2), 185–198.

Parush, A., Kramer, C., Foster-Hunt, T., Momtahan, K., Hunter, A., & Sohmer, B. (2011). Communication and team situation awareness in the OR: Implications for augmentative information display. *Journal of Biomedical Informatics*, *44*(3), 477–485. https://doi.org/10.1016/j.jbi.2010.04.002

Patel, V. M., Warren, O., Humphris, P., Ahmed, K., Ashrafian, H., Rao, C., Athanasiou, T., & Darzi, A. (2010). What does leadership in surgery entail?: Leadership in surgery. *ANZ Journal of Surgery*, *80*(12), 876–883. https://doi.org/10.1111/j.1445-2197.2010.05530.x

Pena, G., Altree, M., Field, J., Sainsbury, D., Babidge, W., Hewett, P., & Maddern, G. (2015). Nontechnical skills training for the operating room: A prospective study using simulation and didactic workshop. *Surgery*, *158*(1), 300–309. https://doi.org/10.1016/j.surg.2015.02.008

Peng, Y., Anton, N. E., Cha, J., Mizota, T., Hennings, J. M., Stambro, R., Rendina, M. A., Stanton-Maxey, K. J., Stefanidis, D., & Yu, D. (2019). Objective Measures of Communication Behavior Predict Clinical Performance. *Journal of Surgical Education*. https://doi.org/10.1016/j.jsurg.2019.03.017

Philiastides, M. G., & Sajda, P. (2007). EEG-Informed fMRI Reveals Spatiotemporal Characteristics of Perceptual Decision Making. *Journal of Neuroscience*, *27*(48), 13082–13091. https://doi.org/10.1523/JNEUROSCI.3540-07.2007

Pluye, P., Gagnon, M.-P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies*, *46*(4), 529–546. https://doi.org/10.1016/j.ijnurstu.2009.01.009

Pomeranz, B., Macaulay, R. J., Caudill, M. A., Kutz, I., Adam, D., Gordon, D., Kilborn, K. M., Barger, A. C., Shannon, D. C., & Cohen, R. J. (1985). Assessment of autonomic function in humans by heart rate spectral analysis. *American Journal of Physiology-Heart and Circulatory Physiology*, *248*(1), H151–H153.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. *Proceedings of the International Conference on HCI*, *2003*.

Proctor, R. W., & Van Zandt, T. (2008). *Human factors in simple and complex systems*. CRC press.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *The Journal of the Acoustical Society of America*, *101*(4), 2267–2277.

Pumprla, J., Howorka, K., Groves, D., Chester, M., & Nolan, J. (2002). Functional assessment of heart rate variability: Physiological basis and practical applications. *International Journal of Cardiology*, *84*(1), 1–14.

Raheem, S., Ahmed, Y. E., Hussein, A. A., Johnson, A., Cavuoto, L., May, P., Cole, A., Wang, D., Ahmad, B., Hasasneh, A., & Guru, K. A. (2018). Variability and interpretation of communication taxonomy during robot-assisted surgery: Do we all speak the same language? *BJU International*, *122*(1), 99–105. https://doi.org/10.1111/bju.14150

Raison, N., Wood, T., Brunckhorst, O., Abe, T., Ross, T., Challacombe, B., Khan, M. S., Novara, G., Buffi, N., Van Der Poel, H., McIlhenny, C., Dasgupta, P., & Ahmed, K. (2017). Development and validation of a tool for non-technical skills evaluation in robotic surgery—The ICARS system. *Surgical Endoscopy*, *31*(12), 5403–5410. https://doi.org/10.1007/s00464-017-5622-x

Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: A review. *Medical and Biological Engineering and Computing*, *44*(12), 1031–1051. https://doi.org/10.1007/s11517-006-0119-0

Rehim, S. A., DeMoor, S., Olmsted, R., Dent, D. L., & Parker-Raley, J. (2017). Tools for Assessment of Communication Skills of Hospital Action Teams: A Systematic Review. *Journal of Surgical Education, 74*(2), 341–351. https://doi.org/10.1016/j.jsurg.2016.09.008

Reiley, C. E., Lin, H. C., Yuh, D. D., & Hager, G. D. (2011). Review of methods for objective surgical skill evaluation. *Surgical Endoscopy*, *25*(2), 356–366. https://doi.org/10.1007/s00464-010-1190-z

Richstone, L., Schwartz, M. J., Seideman, C., Cadeddu, J., Marshall, S., & Kavoussi, L. R. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, *252*(1), 177–182.

Riley, W., Davis, S., Miller, K., Hansen, H., Sainfort, F., & Sweet, R. (2011). Didactic and simulation nontechnical skills team training to improve perinatal patient outcomes in a community hospital. *The Joint Commission Journal on Quality and Patient Safety*, *37*(8), 357–364.

Robertson, E. R., Hadi, M., Morgan, L. J., Pickering, S. P., Collins, G., New, S., Griffin, D., McCulloch, P., & Catchpole, K. (2014). Oxford NOTECHS II: A Modified Theatre Team Non-Technical Skills Scoring System. *PLoS ONE*, *9*(3), e90320. https://doi.org/10.1371/journal.pone.0090320

Rosen, M. A., Dietz, A. S., Lee, N., Wang, I.-J., Markowitz, J., Wyskiel, R. M., Yang, T., Priebe, C. E., Sapirstein, A., Gurses, A. P., & Pronovost, P. J. (2018). Sensor-based measurement of critical care nursing workload: Unobtrusive measures of nursing activity complement traditional task and patient level indicators of workload to predict perceived exertion. *PLOS ONE*, *13*(10), e0204819. https://doi.org/10.1371/journal.pone.0204819

Rosen, M. A., Salas, E., Wu, T. S., Silvestri, S., Lazzara, E. H., Lyons, R., Weaver, S. J., & King, H. B. (2008). Promoting teamwork: An event-based approach to simulation-based teamwork training for emergency medicine residents. *Academic Emergency Medicine*, *15*(11), 1190–1198.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events*. Columbia University.

Sadideen, H., Weldon, S.-M., Saadeddin, M., Loon, M., & Kneebone, R. (2016). A Video Analysis of Intra- and Interprofessional Leadership Behaviors Within "The Burns Suite": Identifying Key Leadership Models. *Journal of Surgical Education*, *73*(1), 31–39. https://doi.org/10.1016/j.jsurg.2015.09.011

Salas, E. E., & Fiore, S. M. (2004). *Team cognition: Understanding the factors that drive process and performance.* American Psychological Association.

Salas, E., Rhodenizer, L., & Bowers, C. A. (2000). The Design and Delivery of Crew Resource Management Training: Exploiting Available Resources. *Human Factors*, *42*(3), 490–511. https://doi.org/10.1518/001872000779698196

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics*, *37*(2), 225–238. https://doi.org/10.1016/j.apergo.2005.02.001

Salmon, P., Stanton, N., Walker, G., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500.

Santen, J. van, Mishra, T., & Klabbers, E. (2008). Prosodic Processing. In J. Benesty, M. M. Sondhi, & Y. A. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 471–488). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49127-9_23

Santos, R., Bakero, L., Franco, P., Alves, C., Fragata, I., & Fragata, J. (2012). Characterization of non-technical skills in paediatric cardiac surgery: Communication patterns. *European Journal of Cardio-Thoracic Surgery*, *41*(5), 1005–1012. https://doi.org/10.1093/ejcts/ezs068

Saus, E.-R., Johnsen, B. H., Eid, J., Riisem, P. K., Andersen, R., & Thayer, J. F. (2006). The effect of brief situational awareness training in a police shooting simulator: An experimental study. *Military Psychology*, *18*(sup1), S3–S21.

Saus, E.-R., Johnsen, B. H., Eid, J., & Thayer, J. F. (2012). Who benefits from simulator training: Personality and heart rate variability in relation to situation awareness during navigation training. *Computers in Human Behavior*, *28*(4), 1262–1268. https://doi.org/10.1016/j.chb.2012.02.009

Sauseng, P., Klimesch, W., Schabus, M., & Doppelmayr, M. (2005). Fronto-parietal EEG coherence in theta and upper alpha reflect central executive functions of working memory. *International Journal of Psychophysiology*, *57*(2), 97–103.

Sawyer, T., Laubach, V. A., Hudak, J., Yamamura, K., & Pocrnich, A. (2013). Improvements in teamwork during neonatal resuscitation after interprofessional TeamSTEPPS training. *Neonatal Network*, *32*(1), 26–33.

Schaefer, H., & Helmreich, R. (1993). The operating room management attitudes questionnaire (ORMAQ). *NASA/University of Texas FAA Technical Report. Austin: University of Texas*.

Scherer, S., Weibel, N., Morency, L.-P., & Oviatt, S. (2012). Multimodal prediction of expertise and leadership in learning groups. *Proceedings of the 1st International Workshop on Multimodal Learning Analytics - MLA '12*, 1–8. https://doi.org/10.1145/2389268.2389269

Sevdalis, N., Healey, A. N., & Vincent, C. A. (2007). Distracting communications in the operating theatre. *Journal of Evaluation in Clinical Practice*, *13*(3), 390–394. https://doi.org/10.1111/j.1365-2753.2006.00712.x

Sevdalis, N., Wong, H. W. L., Arora, S., Nagpal, K., Healey, A., Hanna, G. B., & Vincent, C. A. (2012). Quantitative analysis of intraoperative communication in open and laparoscopic surgery. *Surgical Endoscopy*, *26*(10), 2931–2938. https://doi.org/10.1007/s00464-012-2287-3

Sexton, K., Johnson, A., Gotsch, A., Hussein, A. A., Cavuoto, L., & Guru, K. A. (2018). Anticipation, teamwork and cognitive load: Chasing efficiency during robot-assisted surgery. *BMJ Quality & Safety*, *27*(2), 148–154. https://doi.org/10.1136/bmjqs-2017-006701

Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, *5*. https://doi.org/10.3389/fpubh.2017.00258

Siu, J., Maran, N., & Paterson-Brown, S. (2016). Observation of behavioural markers of non-technical skills in the operating room and their relationship to intra-operative incidents. *The Surgeon*, *14*(3), 119–128.

Sodergren, M. H., Orihuela-Espina, F., Froghi, F., Clark, J., Teare, J., Yang, G. Z., & Darzi, A. (2011). Value of orientation training in laparoscopic cholecystectomy. *BJS (British Journal of Surgery)*, *98*(10), 1437–1445. https://doi.org/10.1002/bjs.7546

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, *30*(2), 337–360.

Stefanidis, D. (2010). Optimal acquisition and assessment of proficiency on simulators in surgery. *Surgical Clinics*, *90*(3), 475–489.

Steinemann, S., Berg, B., DiTullio, A., Skinner, A., Terada, K., Anzelon, K., & Ho, H. C. (2012). Assessing teamwork in the trauma bay: Introduction of a modified "NOTECHS" scale for trauma. *The American Journal of Surgery*, *203*(1), 69–75. https://doi.org/10.1016/j.amjsurg.2011.08.004

Stone, J. L., Aveling, E.-L., Frean, M., Shields, M. C., Wright, C., Gino, F., Sundt, T. M., & Singer, S. J. (2017). Effective Leadership of Surgical Teams: A Mixed Methods Study of Surgeon Behaviors and Functions. *The Annals of Thoracic Surgery*, *104*(2), 530–537. https://doi.org/10.1016/j.athoracsur.2017.01.021

Sun, G., Wanyan, X., Wu, X., & Zhuang, D. (2017). The Influence of HUD Information Visual Coding on Pilot's Situational Awareness. *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, *1*, 139–143.

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39. https://doi.org/10.1177/875647939000600106

Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart Rate Variability, Prefrontal Neural Function, and Cognitive Performance: The Neurovisceral Integration Perspective on Self-regulation, Adaptation, and Health. *Annals of Behavioral Medicine*, *37*(2), 141–153. https://doi.org/10.1007/s12160-009-9101-z

Thomas, A., Campwala, Z., Keheila, M., Ruckle, D., Pierce, M., Mattison, B., West, B., Thomas, J., Hogue, P., Abourbih, S., & Baldwin, D. D. (2019). Impact of a Wireless System Upon Verbal Communication in a Simulated Robotic Operating Theater. *Urology*, *123*, 151–156. https://doi.org/10.1016/j.urology.2018.07.059

Tien, G., Atkins, M. S., Zheng, B., & Swindells, C. (2010). Measuring situation awareness of surgeons in laparoscopic training. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*, 149. https://doi.org/10.1145/1743666.1743703

Tien, G., Zheng, B., & Atkins, M. S. (2011). Quantifying Surgeons' Vigilance during Laparoscopic Operations Using Eyegaze Tracking. *Studies in Health Technology and Informatics*, 658–662. https://doi.org/10.3233/978-1-60750-706-2-658

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G.-Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: A systematic review. *Journal of Surgical Research*, *191*(1), 169–178. https://doi.org/10.1016/j.jss.2014.04.032

Tiferes, J., Bisantz, A. M., Bolton, M. L., Higginbotham, D. J., O'Hara, R. P., Wawrzyniak, N. K., Kozlowski, J. D., Ahmad, B., Hussein, A. A., & Guru, K. A. (2016). Multimodal team interactions in Robot-Assisted Surgery. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 518–522. https://doi.org/10.1177/1541931213601118

Tiferes, J., Bisantz, A. M., & Guru, K. A. (2015). Team interaction during surgery: A systematic review of communication coding schemes. *Journal of Surgical Research, 195*(2), 422–432.

Tiferes, J., Hussein, A. A., Bisantz, A., Higginbotham, D. J., Sharif, M., Kozlowski, J., Ahmad, B., O'Hara, R., Wawrzyniak, N., & Guru, K. (2019). Are gestures worth a thousand words? Verbal and nonverbal communication during robot-assisted surgery. *Applied Ergonomics, 78*, 251–262. https://doi.org/10.1016/j.apergo.2018.02.015

Tiferes, J., Hussein, A. A., Bisantz, A., Kozlowski, J. D., Sharif, M. A., Winder, N. M., Ahmad, N., Allers, J., Cavuoto, L., & Guru, K. A. (2016). The Loud Surgeon Behind the Console: Understanding Team Activities During Robot-Assisted Surgery. *Journal of Surgical Education, 73*(3), 504–512. https://doi.org/10.1016/j.jsurg.2015.12.009

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D., Horsley, T., & Weeks, L. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473.

Undre, S., Sevdalis, N., Healey, A. N., Darzi, A., & Vincent, C. A. (2007). Observational teamwork assessment for surgery (OTAS): Refinement and application in urological surgery. *World Journal of Surgery*, *31*(7), 1373–1381.

Unsworth, K., & West, M. (2000). *Teams: The Challenges of Cooperative Work* (SSRN Scholarly Paper ID 2182831). Social Science Research Network. https://papers.ssrn.com/abstract=2182831

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology, 42*(3), 323–342.

Vine, S. J., Masters, R. S. W., McGrath, J. S., Bright, E., & Wilson, M. R. (2012). Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills. *Surgery*, *152*(1), 32–40. https://doi.org/10.1016/j.surg.2012.02.002

Wadhera, R. K., Parker, S. H., Burkhart, H. M., Greason, K. L., Neal, J. R., Levenick, K. M., Wiegmann, D. A., & Sundt, T. M. (2010). Is the "sterile cockpit" concept applicable to cardiovascular surgery critical intervals or critical events? The impact of protocol-driven communication during cardiopulmonary bypass. *The Journal of Thoracic and Cardiovascular Surgery*, *139*(2), 312–319. https://doi.org/10.1016/j.jtcvs.2009.10.048

Watanabe, H., Makino, T., Tokita, Y., Kishi, M., Lee, B., Matsui, H., Shinozaki, H., & Kama, A. (2019). Changes in attitudes of undergraduate students learning interprofessional education in the absence of patient safety modules: Evaluation with a modified T-TAQ instrument. *Journal of Interprofessional Care*, *33*(6), 689–696. https://doi.org/10.1080/13561820.2019.1598951

Weaver, S. J., Rosen, M. A., DiazGranados, D., Lazzara, E. H., Lyons, R., Salas, E., Knych, S. A., McKeever, M., Adler, L., & Barker, M. (2010). Does teamwork improve performance in the operating room? A multilevel evaluation. *The Joint Commission Journal on Quality and Patient Safety*, *36*(3), 133–142.

Weigl, M., Weber, J., Hallett, E., Pfandler, M., Schlenker, B., Becker, A., & Catchpole, K. (2018). Associations of Intraoperative Flow Disruptions and Operating Room Teamwork During Robotic-assisted Radical Prostatectomy. *Urology*, *114*, 105–113. https://doi.org/10.1016/j.urology.2017.11.060

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., & Movellan, J. (2009). Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 2106–2111.

Wilson, M. R., Poolton, J. M., Malhotra, N., Ngo, K., Bright, E., & Masters, R. S. (2011). Development and validation of a surgical workload measure: The surgery task load index (SURG-TLX). *World Journal of Surgery*, *35*(9), 1961.

Wood, T. C., Maqsood, S., Zoutewelle, S., Nanavaty, M. A., & Rajak, S. (2020). Development of the HUman Factors in intraoperative Ophthalmic Emergencies Scoring System (HUFOES) for non-technical skills in cataract surgery. *Eye*, 1–9. https://doi.org/10.1038/s41433-020-0921-1

Wood, T. C., Raison, N., Haldar, S., Brunckhorst, O., McIlhenny, C., Dasgupta, P., & Ahmed, K. (2017). Training Tools for Nontechnical Skills for Surgeons—A Systematic Review. *Journal of Surgical Education*, *74*(4), 548–578. https://doi.org/10.1016/j.jsurg.2016.11.017

Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., & Yu, D. (2019). Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training. *Human Factors*, 0018720819874544.

Yarkoni, T., Braver, T. S., Gray, J. R., & Green, L. (2005). Prefrontal Brain Activity Predicts Temporally Extended Decision-Making Behavior. *Journal of the Experimental Analysis of Behavior*, *84*(3), 537–554. https://doi.org/10.1901/jeab.2005.121-04

Yu, D., Minter, R. M., Armstrong, T. J., Frischknecht, A. C., Green, C., & Kasten, S. J. (2014). Identification of technique variations among microvascular surgeons and cases using hierarchical task analysis. *Ergonomics*, *57*(2), 219–235.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System. *World Journal of Surgery*, *32*(4), 548–556. https://doi.org/10.1007/s00268-007-9320-z

Yule, S., Flin, R., Maran, N., Youngson, G., Mitchell, A., Rowley, D., & Paterson-Brown, S. (2008). Debriefing surgeons on non-technical skills (NOTSS). *Cognition, Technology & Work*, *10*(4), 265–274.

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery*, *139*(2), 140–149. https://doi.org/10.1016/j.surg.2005.06.017

Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. *Medical Education*, *40*(11), 1098–1104. https://doi.org/10.1111/j.1365-2929.2006.02610.x

Zhang, C., Miller, C., Volkman, K., Meza, J., & Jones, K. (2015). Evaluation of the team performance observation tool with targeted behavioral markers in simulation-based interprofessional education. *Journal of Interprofessional Care*, *29*(3), 202–208. https://doi.org/10.3109/13561820.2014.982789

Zheng, B., Swanström, L. L., & Mackenzie, C. L. (2007). A laboratory study on anticipatory movement in laparoscopic surgery: A behavioral indicator for team collaboration. *Surgical Endoscopy*, *21*(6), 935–940.

Zheng, Bin, Tien, G., Atkins, S. M., Swindells, C., Tanin, H., Meneghetti, A., Qayumi, K. A., & Panton, O. N. M. (2011). Surgeon's vigilance in the operating room. *The American Journal of Surgery*, *201*(5), 673–677. https://doi.org/10.1016/j.amjsurg.2011.01.016

APPENDIX A. SCOPING REVIEW SEARCH TERMS

Table A1. Final search queries for each database

Database	Search query
PubMed	(((non-technical[Title/Abstract] OR "non-technical skills"[Title/Abstract] OR nontechnical[Title/Abstract] OR "human factor"[Title/Abstract] OR "human factors"[Title/Abstract] OR communication[Title/Abstract] OR teamwork[Title/Abstract] OR "team work"[Title/Abstract] OR leadership[Title/Abstract] OR "situation awareness"[Title/Abstract] OR "situational awareness"[Title/Abstract] OR vigilance[Title/Abstract] OR monitoring[Title/Abstract] OR "decision making"[Title/Abstract] OR "decision-making"[Title/Abstract]) AND (physiological[Title/Abstract] OR behavioral[Title/Abstract] OR behavior[Title/Abstract] OR behavior[Title/Abstract] OR assess[Title/Abstract] OR evaluation[Title/Abstract] OR objective[Title/Abstract] OR measure[Title/Abstract] OR empirical[Title/Abstract] OR quantitative[Title/Abstract] OR "heart rate"[Title/Abstract] OR "heart rate variability"[Title/Abstract] OR "HRV"[Title/Abstract] OR "ECG"[Title/Abstract] OR "EKG"[Title/Abstract] OR electrocardiography[Title/Abstract] OR "skin conductance"[Title/Abstract] OR "skin conductance level"[Title/Abstract] OR "SCL"[Title/Abstract] OR "electrodermal activity"[Title/Abstract] OR "EDA"[Title/Abstract] OR "galvanic skin response"[Title/Abstract] OR "GSR"[Title/Abstract] OR "blood pressure"[Title/Abstract] OR ocular[Title/Abstract] OR eye-tracking[Title/Abstract] OR "eye tracking"[Title/Abstract] OR "brain measure"[Title/Abstract] OR "brain activity"[Title/Abstract] OR "EEG"[Title/Abstract] OR electroencephalography[Title/Abstract] OR speech[Title/Abstract] OR interaction[Title/Abstract] OR gesture[Title/Abstract] OR "movement"[Title/Abstract])) AND (surgery[Title/Abstract] OR surgical[Title/Abstract] OR operating[Title/Abstract] OR operation[Title/Abstract] OR "operating room"[Title/Abstract] OR "operating rooms"[Title/Abstract] OR "operating theatre"[Title/Abstract] OR "operating theatres"[Title/Abstract])) AND (clinician[Title/Abstract] OR surgeon[Title/Abstract]) OR ((((("social skills"[MeSH Terms] OR "test taking skills"[MeSH Terms] OR "motor skills"[MeSH Terms] OR "clinical competence"[MeSH Terms] OR "emotional intelligence"[MeSH Terms] OR "thinking"[MeSH Terms] OR "professional competence"[MeSH Terms] OR "professionalism"[MeSH Terms] OR "professionalism"[All Fields] OR "interpersonal relations"[MeSH Terms] OR "crisis intervention"[MeSH Terms] OR "authoritarianism"[MeSH Terms] OR "professional practice"[MeSH Terms] OR "delegation, professional"[MeSH Terms] OR "teach-back communication"[MeSH Terms] OR "interprofessional relations"[MeSH Terms] OR "communication"[MeSH Terms] OR "interdisciplinary communication"[MeSH Terms] OR "crew

	resource management, healthcare"[MeSH Terms] OR "leadership"[MeSH Terms] OR "awareness"[MeSH Terms] OR "decision making"[MeSH Terms] OR "psychomotor performance"[MeSH Terms] OR "problem solving"[MeSH Terms] OR "mental processes"[MeSH Terms] OR "ergonomics"[MeSH Terms]) AND ("heart rate"[MeSH Terms] OR "heart rate determination"[MeSH Terms] OR "electrocardiography"[MeSH Terms] OR "galvanic skin response"[MeSH Terms] OR "blood pressure"[MeSH Terms] OR "blood pressure determination"[MeSH Terms] OR "eye movements"[MeSH Terms] OR "saccades"[MeSH Terms] OR "electroencephalography"[MeSH Terms] OR "feedback, sensory"[MeSH Terms] OR "communication methods, total"[MeSH Terms] OR "manual communication"[MeSH Terms])) AND ("operating rooms"[MeSH Terms] OR "general surgery"[MeSH Terms] OR "surgical procedures, operative"[MeSH Terms] OR "specialties, surgical"[MeSH Terms])) AND "surgeons"[MeSH Terms])

PsycINFO	(TI ((non-technical OR “non-technical skills” OR nontechnical OR “human factor” OR “human factors” OR communication OR teamwork OR “team work” OR leadership OR “situation awareness” OR “situational awareness” OR vigilance OR monitoring OR “decision making” OR “decision-making”)) AND TI ((physiological OR behavioral OR behavior OR behavior OR assess OR evaluation OR objective OR measure OR empirical OR quantitative OR “heart rate” OR “heart rate variability” OR “HRV” OR “ECG” OR “EKG” OR electrocardiography OR “skin conductance” OR “skin conductance level” OR “SCL” OR “electrodermal activity” OR “EDA” OR “galvanic skin response” OR “GSR” OR “blood pressure” OR ocular OR eye-tracking OR “eye tracking” OR “brain measure” OR “brain activity” OR “EEG” OR electroencephalography OR speech OR interaction OR gesture OR movement)) AND TI ((surgery OR surgical OR operating OR operation OR “operating room” OR “operating rooms” OR “operating theatre” OR “operating theatres”)) AND TI ((clinician OR surgeon))) OR (AB ((non-technical OR “non-technical skills” OR nontechnical OR “human factor” OR “human factors” OR communication OR teamwork OR “team work” OR leadership OR “situation awareness” OR “situational awareness” OR vigilance OR monitoring OR “decision making” OR “decision-making”)) AND AB ((physiological OR behavioral OR behavior OR behavior OR assess OR evaluation OR objective OR measure OR empirical OR quantitative OR “heart rate” OR “heart rate variability” OR “HRV” OR “ECG” OR “EKG” OR electrocardiography OR “skin conductance” OR “skin conductance level” OR “SCL” OR “electrodermal activity” OR “EDA” OR “galvanic skin response” OR “GSR” OR “blood pressure” OR ocular OR eye-tracking OR “eye tracking” OR “brain measure” OR “brain activity” OR “EEG” OR electroencephalography OR speech OR interaction OR gesture OR movement)) AND AB ((surgery OR surgical OR operating OR operation OR “operating room” OR “operating rooms” OR “operating theatre” OR “operating theatres”)) AND AB ((clinician OR surgeon))) OR (KW ((non-technical OR “non-technical skills” OR nontechnical OR “human factor” OR “human factors” OR communication OR teamwork OR “team work” OR leadership OR “situation awareness” OR “situational awareness” OR vigilance OR monitoring OR “decision making” OR “decision-making”)) AND KW ((physiological OR behavioral OR behavior OR behavior OR assess OR evaluation OR objective OR measure OR empirical OR quantitative OR “heart rate” OR “heart rate variability” OR “HRV” OR “ECG” OR “EKG” OR electrocardiography OR “skin conductance” OR “skin conductance level” OR “SCL” OR “electrodermal activity” OR “EDA” OR “galvanic skin response” OR “GSR” OR “blood pressure” OR ocular OR eye-tracking OR “eye tracking” OR “brain measure” OR “brain activity” OR “EEG” OR electroencephalography OR speech OR interaction OR gesture OR movement)) AND KW ((surgery OR surgical OR operating OR operation OR “operating room” OR “operating rooms” OR “operating theatre” OR “operating theatres”)) AND KW ((clinician OR surgeon))) OR (SU (Human factors measures OR human factors engineering OR communication OR work

	teams OR leadership OR transactional leadership OR transformational leadership OR leadership style OR awareness OR vigilance OR attention OR monitoring OR decision making OR interpersonal interaction) AND SU surgery AND SU ("surgeons or physicians" OR clinician) AND SU (Psychophysiological measures OR behavioral assessment OR behavior OR psychological assessment OR cognitive assessment OR measurement OR empirical methods OR quantitative methods OR heart rate OR heart rate variability OR electrocardiography OR skin resistance OR psychophysiological measures OR galvanic skin response OR blood pressure OR eye fixation OR visual tracking OR electroencephalography OR neurobiological measures OR oral communication OR speech characteristics OR gestures))

Compendex *(Autostemming On; free text terms were used in the controlled term population concept because no synonym controlled term identified)*	(((((((non-technical OR "non-technical skills" OR nontechnical OR "human factor" OR "human factors" OR communication OR teamwork OR "team work" OR leadership OR "situation awareness" OR "situational awareness" OR vigilance OR monitoring OR "decision making" OR decision-making)) WN KY) AND (((surgery OR surgical OR operating OR operation* OR "operating room" OR "operating rooms" OR "operating theatre" OR "operating theatres")) WN KY)) AND (((clinician* OR surgeon*)) WN KY)) AND (((physiological OR behavioral OR behavior OR behaviour OR assess* OR evaluation* OR objective OR measure* OR empirical OR quantitative OR "heart rate" OR "heart rate variability" OR "HRV" OR "ECG" OR "EKG" OR electrocardiography OR "skin conductance" OR "skin conductance level" OR "SCL" OR "electrodermal activity" OR "EDA" OR "galvanic skin response" OR "GSR" OR "blood pressure" OR ocular OR eye-tracking OR "eye tracking" OR "brain measure" OR "brain activity" OR "EEG" OR electroencephalography OR speech OR interaction* OR gesture* OR movement*)) WN KY)) OR (((((((("human factors" OR "decision making" OR "behavioral research" OR "communication")) WN CV)) AND (("surgery") WN CV)) AND ((("physiological models" OR "quantitative analysis" OR "Biomedical signal processing" OR "Electrocardiography" OR "Blood pressure" OR "eye tracking" OR "Tracking (position)" OR "Behavioral research" OR "Electroencephalography" OR "Acoustics" OR "Speech")) WN CV)) AND (((clinician OR surgeon)) WN KY))))

Inspec *(Autostemming On; free text terms were used in the controlled term population concept because no synonym controlled term identified)*	(((((((((non-technical OR "non-technical skills" OR nontechnical OR "human factor" OR "human factors" OR communication OR teamwork OR "team work" OR leadership OR "situation awareness" OR "situational awareness" OR vigilance OR monitoring OR "decision making" OR decision-making)) WN KY) AND (((surgery OR surgical OR operating OR operation* OR "operating room" OR "operating rooms" OR "operating theatre" OR "operating theatres")) WN KY)) AND (((clinician* OR surgeon*)) WN KY)) AND (((physiological OR behavioral OR behavior OR behaviour OR assess* OR evaluation* OR objective OR measure* OR empirical OR quantitative OR "heart rate" OR "heart rate variability" OR "HRV" OR "ECG" OR "EKG" OR electrocardiography OR "skin conductance" OR "skin conductance level" OR "SCL" OR "electrodermal activity" OR "EDA" OR "galvanic skin response" OR "GSR" OR "blood pressure" OR ocular OR eye-tracking OR "eye tracking" OR "brain measure" OR "brain activity" OR "EEG" OR electroencephalography OR speech OR interaction* OR gesture* OR movement*)) WN KY)))) OR (((((((((("human factors" OR "team working" OR "leadership" OR "decision making")) WN CV)) AND (("surgery") WN CV)) AND ((("physiological models" OR "quantitative analysis" OR "electrocardiography" OR "blood pressure measurement" OR "eye tracking" OR "electroencephalography" OR "speech")) WN CV)) AND (((clinician OR surgeon)) WN KY)))))

Scopus	(TITLE-ABS-KEY(non-technical OR {non-technical skills} OR nontechnical OR {human factor} OR {human factors} OR communication OR teamwork OR {team work} OR leadership OR {situation awareness} OR {situational awareness} OR vigilance OR monitoring OR {decision making} OR decision-making) AND TITLE-ABS-KEY(physiological OR behavioral OR behavior OR behavior OR assess OR evaluation OR objective OR measure OR empirical OR quantitative OR {heart rate} OR {heart rate variability} OR {HRV} OR {ECG} OR {EKG} OR electrocardiography OR {skin conductance} OR {skin conductance level} OR {SCL} OR {electrodermal activity} OR {EDA} OR {galvanic skin response} OR {GSR} OR {blood pressure} OR ocular OR eye-tracking OR {eye tracking} OR {brain measure} OR {brain activity} OR {EEG} OR electroencephalography OR speech OR interaction OR gesture OR movement) AND TITLE-ABS-KEY(surgery OR surgical OR operating OR operation OR {operating room} OR {operating rooms} OR {operating theatre} OR {operating theatres}) AND TITLE-ABS-KEY(clinician OR surgeon))

APPENDIX B. SURVEYS

Hospital ..... Trainer name ..... Date .....

Trainee name ..... Operation .....

Category	Category rating*	Element	Element rating*	Feedback on performance and debriefing notes
Situation awareness		Gathering information		
		Understanding information		
		Projecting and anticipating future state		
Decision-making		Considering options		
		Selecting and communicating option		
		Implementing and reviewing decisions		
Communication and teamwork		Exchanging information		
		Establishing a shared understanding		
		Coordinating team activities		
Leadership		Setting and maintaining standards		
		Supporting others		
		Coping with pressure		

* **1** Poor; **2** Marginal; **3** Acceptable; **4** Good; **NA** Not applicable

1 Poor	Performance endangered or potentially endangered patient safety; serious remediation is required
2 Marginal	Performance indicated cause for concern; considerable improvement is needed
3 Acceptable	Performance was of a satisfactory standard but could be improved
4 Good	Performance was of a consistently high standard, enhancing patient safety; it could be used as a positive example for others
NA	Not applicable

Figure B1. NOTSS assessment tool (S. Yule et al., 2008)

Workload Survey

Directions: For each of the questions below, place a marker (X) on the answer of your choice (0.5 intervals).

1. Mental demand

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?

Low — High

2. Physical demand

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?

Low — High

3. Temporal demand

How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Low — High

4. Effort

How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low — High

5. Performance

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance and accomplishing these goals?

Good — Poor

6. Frustration level

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Low — High

7. Task Complexity

How complex was the task?

Low — High

8. Situational stress

How anxious did you feel while performing the task?

Low — High

9. Distractions

How distracting was the environment?

Low — High

Figure B2. Combination of NASA/SURG-TLX

APPENDIX C. ADDITIONAL CHAPTER 4 RESULTS

Table C1: Individual linear correlations with overall NTS score and communication features where surgeon is sender

Feature	Correlation Coefficient	p-value
Circulating_Callout	0.25	0.02
Circulating_Checkback	0.19	0.07
Circulating_ClosedLoop	0.21	0.05
Circulating_Request	0.30	0.00
Circulating_Question	0.09	0.42
Circulating_Confirmation	0.27	0.01
Circulating_Status	0.08	0.48
Circulating_Irrelevant	0.01	0.92
Circulating_Non_verbal	0.26	0.01
Assistant_Callout	-0.06	0.55
Assistant_Checkback	-0.14	0.18
Assistant_ClosedLoop	0.06	0.59
Assistant_Request	0.00	0.98
Assistant_Question	0.05	0.60
Assistant_Confirmation	0.14	0.18
Assistant_Status	-0.15	0.15
Assistant_Irrelevant	-0.14	0.17
Assistant_Non_verbal	0.10	0.34
Technician_Callout	0.20	0.05
Technician_Checkback	0.15	0.15
Technician_ClosedLoop	0.16	0.12
Technician_Request	0.19	0.07
Technician_Question	0.17	0.11
Technician_Confirmation	0.20	0.05
Technician_Status	0.11	0.32
Technician_Irrelevant	-0.03	0.74
Technician_Non_verbal	0.22	0.03
Anesthesiologist_Callout	0.22	0.03
Anesthesiologist_Checkback	0.11	0.29
Anesthesiologist_ClosedLoop	NaN	NaN
Anesthesiologist_Request	0.27	0.01
Anesthesiologist_Question	0.21	0.05
Anesthesiologist_Confirmation	0.10	0.33

Anesthesiologist_Status	0.11	0.30
Anesthesiologist_Irrelevant	0.08	0.43
Anesthesiologist_Non_verbal	0.20	0.06
Team_Callout	0.18	0.08
Team_Request	0.19	0.08
Team_Question	0.17	0.11
Team_Status	0.15	0.14
Team_Irrelevant	0.00	0.97
Team_Non_verbal	0.17	0.10

Table C2: Individual linear correlations with overall NTS score and communication features where surgeon is receiver

Feature	Correlation Coefficient	p-value
Circulating_Callout	0.15	0.17
Circulating_Checkback	0.12	0.26
Circulating_ClosedLoop	0.26	0.01
Circulating_Request	0.17	0.11
Circulating_Question	0.28	0.01
Circulating_Confirmation	0.17	0.11
Circulating_Status	-0.03	0.76
Circulating_Irrelevant	0.04	0.72
Circulating_Non_verbal	0.16	0.14
Assistant_Callout	-0.12	0.25
Assistant_Checkback	-0.12	0.26
Assistant_ClosedLoop	0.03	0.76
Assistant_Request	0.00	0.99
Assistant_Question	-0.06	0.59
Assistant_Confirmation	-0.08	0.43
Assistant_Status	-0.11	0.31
Assistant_Irrelevant	-0.12	0.26
Assistant_Non_verbal	-0.16	0.13
Technician_Callout	-0.04	0.68
Technician_Checkback	0.14	0.18
Technician_ClosedLoop	0.04	0.71
Technician_Request	0.21	0.04
Technician_Question	0.17	0.12
Technician_Confirmation	0.06	0.60
Technician_Status	0.01	0.89
Technician_Irrelevant	0.02	0.87
Technician_Non_verbal	-0.12	0.26
Anesthesiologist_Callout	0.11	0.31
Anesthesiologist_Checkback	0.16	0.12
Anesthesiologist_ClosedLoop	NaN	NaN
Anesthesiologist_Question	0.11	0.31
Anesthesiologist_Confirmation	0.24	0.02
Anesthesiologist_Status	0.14	0.17
Anesthesiologist_Irrelevant	0.08	0.44
Anesthesiologist_Non_verbal	0.15	0.16

Table C3: Individual linear correlations with overall NTS score and speech features

Feature	Correlation Coefficient	p-value
speech_dur_mean	0.43	0.00
intensity_mean	0.04	0.69
Pitch_mean	0.40	0.00
articulate_rate_mean	0.10	0.35
dur_diff_mean	-0.19	0.08
int_diff_mean	-0.13	0.25
pit_diff_mean	-0.06	0.58
art_rate_diff_mean	0.13	0.27
duration_burst	0.21	0.06
intensity_burst	-0.03	0.76
pitch_burst	0.21	0.06
art_rate_burst	-0.17	0.13

Table C4: Individual linear correlations with overall NTS score and proximity features

Feature	Correlation Coefficient	p-value
Assisting1PercentClose	0.44	0.03
Assisting1PercentNear	0.43	0.04
Assisting1PercentFar	-0.22	0.30
Anesthesiologist1PercentClose	-0.06	0.74
Anesthesiologist1PercentNear	-0.20	0.31
Anesthesiologist1PercentFar	0.13	0.52
Circulating1PercentClose	-0.07	0.70
Circulating1PercentNear	0.43	0.01
Circulating1PercentFar	-0.45	0.01
ScrubTech1PercentClose	-0.06	0.74
ScrubTech1PercentNear	-0.05	0.75
ScrubTech1PercentFar	0.18	0.29
PatientBedPercentClose	-0.33	0.38
PatientBedPercentNear	-0.44	0.23
PatientBedPercentFar	-0.22	0.57
Assisting1OverallMean	-0.25	0.25
Assisting1CloseMean	0.33	0.12
Assisting1NearMean	0.26	0.21
Assisting1FarMean	-0.15	0.47
Anesthesiologist1OverallMean	-0.02	0.91
Anesthesiologist1CloseMean	0.00	0.99
Anesthesiologist1NearMean	-0.17	0.40
Anesthesiologist1FarMean	-0.07	0.71
Circulating1OverallMean	-0.34	0.04
Circulating1CloseMean	-0.17	0.33
Circulating1NearMean	0.18	0.29
Circulating1FarMean	-0.29	0.08
ScrubTech1OverallMean	0.26	0.13
ScrubTech1CloseMean	-0.17	0.31
ScrubTech1NearMean	-0.16	0.35
ScrubTech1FarMean	0.24	0.15
PatientBedOverallMean	-0.48	0.19
PatientBedCloseMean	-0.49	0.18
PatientBedNearMean	-0.26	0.48
PatientBedFarMean	-0.29	0.45
Assisting1OverallSD	0.11	0.60
Assisting1CloseSD	0.10	0.66
Assisting1NearSD	0.26	0.22

Assisting1FarSD	0.06	0.80
Anesthesiologist1OverallSD	-0.28	0.15
Anesthesiologist1CloseSD	-0.08	0.70
Anesthesiologist1NearSD	-0.35	0.06
Anesthesiologist1FarSD	-0.17	0.38
Circulating1OverallSD	-0.26	0.13
Circulating1CloseSD	-0.16	0.34
Circulating1NearSD	-0.07	0.70
Circulating1FarSD	-0.14	0.40
ScrubTech1OverallSD	-0.07	0.68
ScrubTech1CloseSD	-0.03	0.86
ScrubTech1NearSD	-0.13	0.44
ScrubTech1FarSD	-0.06	0.74
PatientBedOverallSD	-0.42	0.26
PatientBedCloseSD	-0.41	0.27
PatientBedNearSD	-0.31	0.39
PatientBedFarSD	-0.12	0.76