AFRL-RH-WP-TR-2022-0100

FIDELITY OF MEDICAL SIMULATION FOR CLINICAL SKILL ACQUISITION

James P. Bliss & Kent Etherton
CAE USA

October 2022

Final Report

Distribution A. Approved for public release: distribution unlimited.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
HUMAN EFFECTIVENESS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2022-0100 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
TEVIN J MILLER, 2d Lt, USAF
Work Unit Manager
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signature//
JOHN L. CAMP, DR-III, Ph.D.
Operation Learning Sciences Branch
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signature//
LOUISE A. CARTER, DR-IV, Ph.D.
Chief, Warfighter Interactions and Readiness Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

Michael Hall
Senior Manager, Contracts
T. 817-619-2057
E. michael.hall@caemilusa.com

7 April 2023

In Reply Refer to: WRT-358

USAF/AFMC
AFRL/RAKHA
2310 Eighth Street, Building 45
Wright-Patterson AFB, OH 45433-7541

Attention: Stephanie Auld, Contracting Officer

Subject: Task Order (TO) 0019, Fidelity of Medical Simulation for Clinical Skill Acquisition – Final Report Clarification

Reference: (a) Contract FA8650-16-D-6582, Order FA8650-20-F-6216: TO 0019 Fidelity of Medical Simulation for Clinical Skill Acquisition
(b) Task Order (TO) 0019, *Fidelity of Medical Simulation for Clinical Skill Acquisition – Final Report*, dated November 2022

Dear Stephanie,

As it relates to the reference (b) final report for Task Order 0019 submitted by CAE USA Inc. ("CAE"), this report is marked with Distribution Statement B per the contractual requirements as stated in the reference (a) contract and task order. Despite the Distribution Statement B, including the associated "Proprietary Information" language marking, this report does not contain any CAE Proprietary Information, and as it relates to CAE Proprietary Information, CAE concurs with making this report available for public release.

If you have any additional questions or concerns, please feel free to contact the undersigned at +1 817-619-2057 or by e-mail at michael.hall@caemilusa.com.

Sincerely,

Michael Hall
Senior Manager, Contracts

CAE USA Inc. | 2200 Arlington Downs Road, Arlington, TX 76011 | 817-619-2000 | www.cae.com/cae-usa

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED: START DATE	3. DATES COVERED: END DATE
October 2022	Interim	1 May 2020	1 October 2022

3. TITLE AND SUBTITLE

Fidelity of Medical Simulation for Clinical Skill Acquisition

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8650-20-F-6216		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		H117

6. AUTHOR(S)

Dr. James Bliss / Dr. Kent Etherton

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

CAE Inc. / AFRL 711th HPW RHWL
2620 Q Street
Wright-Patterson AFB, OH 45433

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Human Effectiveness Directorate
Warfighter Interactions & Readiness Division
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'SREPORT NUMBER(S)

AFRL-RH-WP-TR-2022-0100

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

AFRL-2023-4738; Cleared 25 Sep 2023

14. ABSTRACT

The scope of this work was to identify, analyze, and develop technologies and methods to study and address the acquisition, retention, and relearning of skills in the context of DoD medical training. Work was performed with the intent to addresses clinical knowledge and skill requirements across a range of military medical operations, including those outlined within the curricular guidelines for Tactical Combat Casualty Care (TCCC), the Comprehensive Medical Readiness Program (CMRP) and Individual Critical Task Lists (ICTL). Investigators leveraged learning sciences to evaluate learning by generating experimental designs and incorporating clinical training events. Also, this effort includes the development of a fidelity taxonomy and evaluation method to inform questions about the curriculum selection (what), resources (how), and timing (when) of optimizing learning and training.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
U	U	U	SAR	124

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Lt Tevin Miller	

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

In recent years, the U.S. Air Force (USAF) has become increasingly reliant on simulator technology to train combat medical personnel. Human patient simulators (HPS) can be a useful tool for replicating procedural elements, but they offer varying amounts of fidelity to represent important task elements. Furthermore, medics often train in groups with HPSs and may receive varying amounts of actual hands-on training. Consequently, defining and objectively quantifying readiness standards for medical Air Force Specialty Codes (AFSC) and individual trainees is difficult and trainee competence and confidence may be confused. The research described within this document was designed to address this by investigating the relationship between simulator capabilities and the ability to achieve personalized learning of combat medical personnel. This was achieved by determining the functional capability of two human performance simulators to train surgical cricothyroidotomy, obtaining expert ratings of simulator effectiveness, and collecting performance data from experienced and inexperienced participants as they used the simulators. The research team completed three phases of data collection: (1) compilation of expert opinions to supplement completion of a task analysis (Appendix A); (2) behavioral training data in a laboratory environment; (3) self-report perceptions data from training experts in the field, completion of a literature review on cricothyroidotomy simulation (Appendix B), and completion of a missing simulator cue analysis (Appendix C). Analyses yielded important conclusions about the importance of cognitive task analysis, triangulation of diverse performance measures, distinction of learner confidence and competence, assessment of training transfer, and consideration of the combined roles of didactic and simulation based practical instruction. In the short term, benefits of the research include resources to standardize training using high-resolution human and system measurement, guidance for simulator procurement, guidance for simulator training, and potential to tailor personalized learning for medics and other AFSCs. More strategic benefits include establishment of baseline data from which to document training gaps and evaluate the impact of future simulator related and curricular changes.

The research capabilities developed, integrated, and implemented during this task order are intended to form the foundation of a streamlined approach for procuring and implementing trauma care simulation technology for training medical personnel. Existing approaches for achieving that goal have focused on capabilities of simulation platforms (Ruisi, 2017) and anticipated cost of technology purchase and implementation (Rooney et al., 2018). The research team's activities intended to supplement such methods with a more detailed examination of the concept of simulation fidelity.

This final report describes the work completed as part of the Warfighter Readiness and Training (WRT) Program, Task Order 0019 (TO-19): Fidelity of Medical Simulation for Clinical Skill Acquisition, from May 2020 through October 2022. The scope of this work was to identify, analyze, and develop technologies and methods to study and address the acquisition, retention, and relearning of skills in the context of Department of Defense (DoD) medical training. Work was performed with the intent to address clinical knowledge and skill requirements across a range of military medical operations, including those outlined within the curricular guidelines for Tactical Combat Casualty Care (TCCC), the Comprehensive Medical Readiness Program (CMRP) and Individual Critical Task Lists (ICTL). Investigators intended to leverage learning sciences to evaluate individual and team learning by generating experimental designs and incorporating clinical training events. Also, this effort includes the development of a fidelity taxonomy and evaluation method to inform questions about the curriculum selection (*what*), resources (*how*), and timing (*when*) of optimizing learning and training.

Overall, the efforts on the current task order resulted in the following major accomplishments:

- Development and execution of a fidelity evaluation method
  - Cricothyroidotomy cognitive task analysis (Appendix A)
  - Cricothyroidotomy literature review (Appendix B)
  - Laerdal SimMan ALS/SynDaver Adult Cricothyroidotomy Trainer Missing Simulator Cue Analysis (Appendix C)
  - Training performance data collection and analysis
  - Fidelity perceptions data collection and analysis
- Utilization of confidence-competence calibration analyses in surgical training environment
- Development of a minimal data abbreviated version of the fidelity evaluation method (Appendix D)
- Presentations and publications
  - Manuscript on cricothyroidotomy skill acquisition published in the Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care (Bliss, Etherton, Hodge, & Winner, 2022a)
  - Manuscript synthesizing various models of fidelity within a novel validity-focused theoretical model (in preparation; Etherton, Bliss, & Winner, 2023)
  - Manuscript leveraging analyses of miscalibrated confidence in trainees (under review; Etherton, Bliss, & Winner, 2022)
  - Poster on surgical error analysis presented at Human Factors and Ergonomics Society Annual Conference (Bliss et al., 2021)
  - Poster on eye tracking metrics during surgical training at DoD Human Factors and Ergonomics Technical Advisory Group (DoDHFETAG; Bliss, Etherton, Hodge, & Winner, 2022b)

2.0 INTRODUCTION

2.1 Statement of the Problem

Increasingly, civilian medical training has relied on human patient simulators as a tool to complement classroom training for medical personnel. The U.S. Air Force has followed suit, embracing part- and whole-task simulation devices to supplement academic training for medically trained personnel (Dorlac, Bishop, & Dorlac, 2014). Such devices, however, vary widely in their ability to faithfully represent the elements of medical procedures (and therefore the procedures *in toto*). Though their appearance may suggest an ability to present appropriate cues in a way that fosters learning, physical functionality frequently lacks veridicality (Winner & Millwater, 2019). Compounding the problem, no comprehensive guide exists to document the fidelity of commonly used human patient simulators relative to each other or to medical procedure requirements. As a result, procurement and adoption decisions often rely on word of mouth, manufacturer claims, or perceived affordability concerns. Unfortunately, such influences jeopardize the ability to complement classroom training with effective and faithful procedural practice. Therefore, trainees may adopt ineffective or harmful methods, putting patients at risk.

Researchers and practitioners are interested in fidelity evaluation because fidelity is a critical predictor of training efficacy. The value of a simulator depends on how faithfully it represents elements of the real-world task such that training on the simulator reliably leads to transfer of skills. However, evaluators have identified that greater fidelity does not always lead to better performance (Carey & Rossler, 2022). That is, at a certain point, simulator developers see diminishing returns with training efficacy compared to the incremental cost of improving simulator realism. Thus, there is considerable debate regarding the “sweet spot” of fidelity that maximizes fidelity while minimizing cost. The purpose when evaluating a simulator’s fidelity is to select which simulators represent the best approximation of the real-world task that can be used for training. As such, we believe evaluators might consider reevaluating fidelity and consider the concept through the lens of validation.

Thus, we argue fidelity evaluation’s value simply lies in determining a simulator’s ability to predict real-world task performance. Regardless of how one defines the term ‘fidelity’, the goal is to determine how predictive performance on the simulator is of performance on the real-world task. There is extensive confusion in the literature regarding how one might define ‘fidelity’; such confusion leads to inconsistency in the way people invoke the term. Research suggests evaluators would benefit from simplifying their focus to addressing the validity of simulator performance metrics relative to the real-world task. Such a comparison would render fidelity evaluation to be a form of validity testing.

2.2 Goal of the Research

Criterion definitions of readiness have been explored widely within the scientific, military and AFRL communities. Our intent was to identify novel approaches for determining readiness in the context of a particular, high-consequence, combat care task: surgical cricothyroidotomy. That task was selected for several reasons. First, it is representative of the “Airway” category of procedures within the TCCC Massive Hemorrhage, Airway, Respirations, Circulation, Head Injury/Hypothermia (MARCH)- Pain, Antibiotics, Wounds, Splinting (PAWS) group. Therefore, TCCC continually publishes updated guidelines for training trauma care specialists to perform the task. Second, considerable research has been conducted concerning surgical cricothyroidotomy. Included within those efforts have been task analyses that prescribe the procedural components necessary for successful performance (c.f., Demirel et al., 2016). Third, numerous simulator comparisons documented in the scientific literature have included surgical cricothyroidotomy as a task of interest. Because of this, there exists a wealth of prior results that bear relevance to the current investigation. Fourth, the task itself is straightforward; criteria for success are clear and errors are generally easy to document. Fifth, simulation training devices used for training surgical cricothyroidotomy are numerous and varied. Civilian and military instructors regularly use full-body manikins, part-task trainers, animal models, three-dimensional (3D)-printed synthetic skin, immersive trainers, and cadavers to train the task; however, rigorously estimating physical and functional fidelity of these options has been attempted by only a few researchers (see Pandian et al., 2020; Mabry, Nichols, Shiner, Bolleter, & Frankfurt, 2014, for exceptions).

Determining readiness to perform surgical cricothyroidotomy after training has traditionally required demonstration of successful task completion on a simulator within a prescribed time period (usually two minutes). However, the premise of the investigation was that simulators likely vary regarding physical and functional realism. For that reason, it was decided to comprehensively measure performance speed and accuracy as well as self-efficacy, situation awareness, workload, attention, and efficiency.

The Air Force Research Laboratory (AFRL) has been seen as uniquely qualified to address this area of research for many reasons. First, AFRL has unique research capability and experience in medical simulation. Wright-Patterson Air Force Base (WPAFB), for example, is the home to the 88th Medical Design Group as well as the USAF School of Aerospace Medicine (USAFSAM). Both are users of simulation based medical training devices and are therefore invested in their effectiveness. Also, AFRL has long been a supporter of the Combat Air Force's (CAF) Live, Virtual, Constructive (LVC) vision of training. Simulation devices for medicine make use of live (standardized patients), virtual (immersive training devices), and constructive (incorporated task scenarios and performance management elements within trainers) features. AFRL is also instrumental in supporting current and new core function master plans in major areas related to medical simulator research and training.

The use of fidelity to categorize and select simulation devices for military medical training is a well-established goal within AFRL (Holt et al., 2012; Prost et al., 2007, 2008) in the early 2010's. Contractor and government scientists worked together to develop a model and strategies for optimizing fidelity levels for less-than-full fidelity simulation systems. With the Deployable Tactical Trainer (DTT) as a model system and process, scientists established a reproducible method that stressed summative and formative assessments of simulator training capabilities, impacts of simulator enhancements, and operator use. Central to their method was a fidelity evaluation methodology that stressed definition of scope, training capability assessments, deficiency surveys, and solution surveys.

TO-19 was intended to extend the fidelity assessment philosophy and methodology beyond its application to fast jet training simulation platforms. In doing so, the research team emphasized several theoretical and applied considerations to guide their efforts:

- Fidelity research conducted over the prior 15 years (from 2004 to the present)
- Unique fidelity demands of tactical combat casualty care for military medical trainees

- Psychometric enhancements to improve the precision of fidelity evaluation
- Demographic and expertise influences on fidelity assessments, especially related to learner confidence and competence
- Incorporation of performance and self-report data that may be sensitive to simulation fidelity differences

2.3 Theoretical Background

Many theoretical frameworks exist to characterize the relation between training device fidelity and associated acquisition and retention of skills. Clearly, the inclusive work of Hays and Singer (1989) and Roza (2004) has been instrumental to specify variations and impacts of training device fidelity. Additional relevant resources by category are described below.

Early investigations of simulator fidelity by AFRL scientists occurred in the 1950s. Miller (1954) at WPAFB distinguished "engineering fidelity" from "psychological fidelity." Subsequent authors equated these terms with "physical" and "functional" fidelity, respectively (Maran & Glavin, 2003). The degree to which a simulator faithfully allows practice on operational tasks is clearly of value for medical training. Complicating conceptualizations of functional fidelity, however, is the acknowledgment of the importance of individual trainee perception. As noted by Tun, Alinier, Tang, and Kneebone (2015), cues that are or are not represented faithfully by the training simulation embody critical influences that facilitate or degrade acquisition of skill. Considered in this way, assessments of "functional" fidelity should necessarily include the perceptual (visual, auditory, tactile) cues required for task understanding and execution. Roza (2004), in his extended discussion of simulation fidelity, acknowledged related complications. However, his explanation stressed the importance of abstract notions of reality associated with broader networked simulations.

Pervasive challenges associated with fidelity of training simulations have plagued the military medical training community. Researchers and practitioners alike continue to conceptualize training technologies as "high fidelity" with minimal empirical justification. Procurement officers tend to value the concept of fidelity, but frequently associate it with functionality. Consequently, training simulators are acquired and used based on the capabilities they support, but without sufficient consideration for how realistically they mirror real tasks based on sensory cues available to operator trainees. Compounding the problem, assessment of simulator fidelity is a time-consuming process that relies on objective and self-report data. Bell and Waag's (1998) recommendations for flight simulator evaluation prioritize analyzing specific simulator characteristics. Also, they emphasize the importance of transfer of training results (to alternative simulation, flight, and combat environments), along with utility evaluation and quantification of performance improvement. Winner and Millwater (2019) demonstrated the potential for applying Bell and Waag's recommendations to combat-medical procedure training.

Investigators within AFRL RHWA (now RHWL) have a lengthy history of investigating simulation fidelity as a part of the WRT Contract. Specific efforts have targeted fast jet (F-16) simulation platforms, Joint Terminal Attack Controller (JTAC) simulation systems and trauma-based combat care simulators. TO 19 was intended as a continuation and extension of those historic efforts.

Further, we believe it is important to nest the discussion of fidelity evaluation within a framework of supporting validity inferences. The reasoning for doing so is because we believe fidelity is only valuable to the simulator procurement and/or evaluation community insofar as it informs the validity of the simulator as a predictor of future real-world performance. As such, we used existing theories of validation (Binning & Barrett, 1989) to guide the efforts of the current task order.

Binning and Barrett (1989) describe a model of validation. In their model, the authors describe how researchers largely agree "validity is not a characteristic of a test or assessment procedure but, instead, of inferences made from test or assessment information" (p. 478; Cronbach, 1970; Guion, 1980, 1987; Landy, 1986; Society for Industrial and Organizational Psychology, 1987; American Psychological Association, 1985; as referenced by Binning & Barrett, 1989). In their model, Binning and Barrett (1989) identified multiple key components relevant to the process by which one provides evidence for validity. Such components include observed measures, latent constructs, and the interrelations between them that constitute various validity inferences. Importantly, they described how actual performance can be defined (through job or task analysis) as a performance domain consisting of a series of performance-relevant behaviors and outcomes. Further, the performance domain can be operationalized as observable metrics (e.g., completion time, error rate. The observed metrics may be significantly related to each other or not. The predictor measure may relate to the criterion performance domain. Finally, the observable predictor and criterion measures may relate to common underlying latent factors that may explain relationships between said measures. For more detail on the original conceptualization of these inferences and how they may be supported with empirical data, see Binning and Barrett (1998).

2.4 Foundations of Physical Fidelity

One of the earliest treatments of simulation fidelity was offered by Gerathewohl (1969). Supplemented later by Rehmann et al. (1995), it offered distinctions among the concepts of physical, functional, psychological, behavioral, engineering, visual, and auditory fidelity. Since that time, there have been other notable reports and articles associating aspects of fidelity with learning effectiveness (c.f., Bredmose, Habig, Davies, Grier, & Lockey, 2010; Lapkin & Levett-Jones, 2011). Most accepted definitions of fidelity have emphasized the degree of similarity between a simulated experience and the actual (operational or real-world) experience. As simulation has continued to be embraced as a training tool by the military medical community, considerations of fidelity have hinged around physical simulator aspects that reflect the advancement of technology. Such considerations are naturally important for impressions of face and content validity; however, functionality is clearly of greater importance for training utility.

2.5 Characterization and Varieties of Functional Fidelity

Miller (1954) at WPAFB distinguished "engineering fidelity" from "psychological fidelity." Subsequent authors equated these terms with "physical" and "functional" fidelity, respectively (Maran & Glavin, 2003). The degree to which a simulator faithfully allows practice on operational tasks is clearly of value for medical training. Complicating conceptualizations of functional fidelity, however, is the acknowledgment of the importance of individual trainee perception. As noted by Tun, Alinier, Tang, and Kneebone (2015), cues that are or are not represented faithfully by the training simulation embody critical influences that facilitate or degrade acquisition of skill. Considered in this way, assessments of "functional" fidelity should necessarily include the perceptual (visual, auditory, tactile) cues required for task understanding and execution. Roza (2004), in his extended discussion of simulation fidelity, acknowledged related complications. However, his explanation stressed the importance of abstract notions of reality associated with broader networked simulations.

2.6 Development of Cognitive Scripts to Facilitate Complex Procedural Learning

One of the most important contributions to modern understanding of skill learning is the work of Schmidt and Bjork (1989, 1992). The authors describe task learning as a process whereby an individual develops some "relatively permanent capability for responding" (p. 208). Conclusions drawn by these authors suggest that cue presentation should be available during learning, particularly as it supports manipulation of difficulty and variability. The idea is that experiencing difficult and variable practice may lead to cognitive scripts that are more robust to decay. Certainly, the characterization and definition of task difficulty is important to refine as a function of subject matter expert input.

The notion of warfighters building cognitive scripts reflects a general notion of "cognitive readiness" advocated in the early 2000s by the Office of the Deputy Undersecretary of Defense for Science and Technology. The term referred to achieving an intuitive understanding of battlefield elements to enable superior pattern recognition, creative adaptability, and intuitive decision making. Such skills are also central for combat medical personnel as they perform medical duties (Fautua & Schatz, 2012).

2.7 Specifications of Training Strategies to Ensure Maximal Learning and Transfer

Comprehensive treatments of the subject have been offered by Adams (1987), Holding (1965), and Osgood (1949) among others. Because of the unique and complex nature of critical medical tasks, researchers have published recommendations specific for medical task mastery. As noted by Scalese, Obeso, and Issenberg (2007), Computer-enhanced mannequins (CEM) have held a central role as part-task medical trainers because of their ability to reproduce the anatomy of a procedure as well as normal and pathophysiological functions. Most recently, virtual reality simulations have allowed replication of environmental aspects for collective training. Since the introduction of the SimOne in 1967, CEM's have shown promise for training. However, Scalese et al. mention several challenges associated with CEM employment: engineering limitations, psychometric requirements and prohibitive cost. Historically, Miller's (1990) "knowledge-applied knowledge-performance" or "see one-do one-teach one" philosophy of medical competency has been a standard practice. If the challenges Scalese et al. mention can be overcome, CEM's may effectively augment that practice.

2.8 Refinement and Adoption of Simulators for Medical Use

Researchers and practitioners have demonstrated the value of computer enabled simulation devices for decades, citing rapid skill acquisition, extended skill retention, and advantages of cost and environmental impact. A particularly valuable resource for documenting advantages and challenges of simulators for military use was offered by Vreuls and Obermayer (1985). They discussed the need for automated measurement of human performance within simulations that is quantifiable and related to task structure. In medical task settings, the demand for powerful performance measurement exists alongside the requirements for physical and functional fidelity of the simulator. If a simulator does not include appropriate sensory cues, it will fail to elicit necessary task performances and will therefore lack utility across the operational spectrum.

Considering scope and findings of prior fidelity research, there are several important areas that call for research:

- Defining cues and tasks that influence physical and functional fidelity within military medical simulations.
- Documenting absolute and relative levels of fidelity associated with specific human patient simulation platforms.
- Establishing and tracking performance indicators to support fidelity ratings of human patient simulation platforms.
- Sustaining skills established during training sessions with human patient simulations.

Considered together, these goals align well with the core research areas associated with AFRL's Human Performance Wing (HPW) "Research-Human" (RH) Area. For example, cognitive neuroscience foci within RH specify the importance of validating assessments of cognitive state and readiness and predicting performance. There is also an emphasis on multisensory cues within Multisensory Perception and Communication; this appears relevant to the cue approach that underlies the proposed research. The multi-phase approach adopted for this research was intended to address the research needs listed above. In Phase 1, we conducted a cognitive task analysis to determine the physical capability of two specific human patient simulators to faithfully support airway maintenance skill practice. In Phase 2, we gathered performance and self-report data to support the judgments from Phase 1. In Phase 3, we extended our analyses to include expert evaluations of simulators used in the field.

2.9 Documentation of Fidelity Associated with Common Human Patient Simulation Devices

As noted by several researchers, the adoption and growth of medical simulation technology has been associated with increased concerns about the potential of simulators to faithfully represent the functional and perceptual elements of the target procedure and of human systems (see Krishnan, Keloth, & Ubedulla, 2017). It is increasingly common to see clinical educators integrate CEM's into training curricula. The assumption voiced by many is that such manikins are "high fidelity" because they represent physical and functional elements with some precision. However, comprehensive examinations of fidelity are scarce. Many evaluate fidelity as it relates to trainee impressions of broad constructs such as confidence or satisfaction (c.f., Crouch, 2010). Other popular descriptors include realism (Feingold, Calaluce, & Kallen, 2004) and effectiveness (Mayo, Hackney, Muech, Ribaudo, & Schneider, 2004).

With few exceptions, impressions of simulator fidelity are obtained from entry-level medical students (c.f., Rhodes & Curran, 2005). However, as Rodgers (2007) accurately points out, CEM's may hold value for learners at all levels. Also, it is possible that entry-level medical students may lack the knowledge to fully evaluate the fidelity of a training device. This is particularly the case if the object of evaluation is not just the CEM itself, but the functionality of the CEM in conjunction with the broader task performance environment.

An important but often overlooked aspect of evaluation concerns the specific tasks that are the targets of simulation. In many cases, the choice of tasks reflects only those that are specifically targeted during instruction. Though this approach is reasonable, it may not be comprehensive. A preferable alternative is to conduct a functional and hierarchical task analysis. By doing so, capabilities and limitations of the simulation device may be documented according to knowledge, skill, ability or other characteristics. It may also be possible to determine whether categories of tasks are able to be faithfully represented by the CEM. The results of such an analysis may be of help to instructors, students, simulation designers, and procurement personnel.

Given the importance of associating medical simulation capabilities with targeted tasks for training, our first year of research was focused on documenting the fidelity aspects of a variety of commonly used medical human patient simulators. We performed a task analytic investigation to specify the tasks that are formally associated with major procedures included in the MARCH-PAWS taxonomy (Appendix A). The investigation was partially dependent on established resources and completed task analyses that were available. It also reflected subject matter expert endorsement of the task structure we documented. Because of the importance of physical and functional cues, the task analysis was separated into cognitive and physical components. Our emphasis on simulator functions required us to carefully document tasks in a functional as well as a hierarchical fashion.

Once the task analysis was completed, we selected specific CEM's to be the targets of our analyses. Initial candidate systems included the following:

- CAE Apollo – As the latest iteration of the METI product, CAE's adult male patient simulator that integrates with CAE learning modules. It offers pre-programmed scenarios for anaphylaxis, heart failure with pulmonary edema, severe young asthmatic, and subdural hematoma.
- CAE Caesar – This platform is specifically designed for point-of-care medical treatment of extreme trauma cases. It offers ten scenarios relevant for military and civilian operations.
- ITTS Tactical Operations Medical (TOM)/TAMI Manikins – These platforms are also used to represent tactical combat casualty care. They offer the ability to build scenarios. They are wirelessly reactive and can represent gunshot wounds, blast injuries and burns.
- Laerdal SimMan 3G – The SimMan3G is a comprehensive CEM that is responsive to 145 drugs. It has an extensive library of clinical case scenarios and may be expanded with add-on technologies.
- Laerdal SimMan ALS – Advertised as a mobile system, this product is particularly appropriate for training basic assessment and advanced life-support skills. It also may be combined with other products to extend its applicability for a variety of conditions.
- SynDaver Adult Cricothyroidotomy Trainer – This trainer features advanced artificial tissue representation to enable realistic tactile sensations as individuals perform surgical tasks. This particular system is a part-task trainer specifically designed for use to train airway procedures such as surgical cricothyroidotomy.

Data collection has long been an essential prerequisite for understanding and addressing performance constraints in combat care environments. To accurately and comprehensively determine the extent to which critical skills may be acquired, retained, refreshed and transferred to realistic environments and conditions, researchers have emphasized the importance of conducting training research that incorporates measurement of conventional performance indicators such as procedure completion time, skill efficiency, errors made, and task knowledge (Vreuls & Obermayer, 1985). More recently, researchers have exploited quantitative and technological advancements such as more powerful tools for statistical analysis, hardware, and software solutions for measurement of physiological indicators of cardiac, muscular, and neuronal activation, and behavioral tracking methods such as eye tracking. Though such advancements come with psychometric challenges related to measurement validity, reliability, and sensitivity, they have offered assessments that are accomplished more rapidly and with comparatively less error. Such advantages have been particularly beneficial for the measurement of cognitive constructs such as attention, workload, stress, and resilience.

3.0 OVERALL METHOD

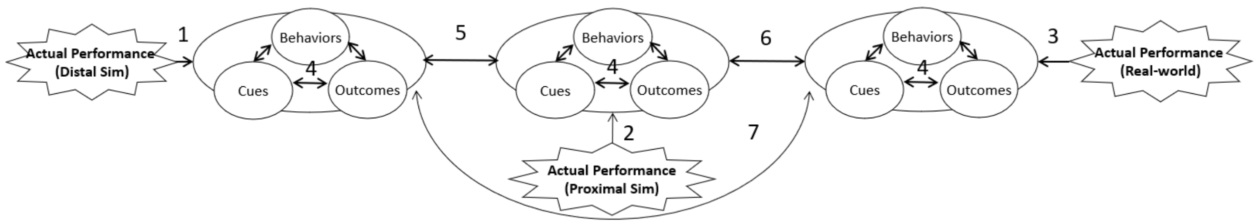
The contractor team developed a simulator fidelity evaluation method (Table 1). This method included seven steps meant to provide a more comprehensive evaluation of simulation fidelity using a combination of qualitative and quantitative insights. First, it was important to identify the relevant validity inferences involved in the simulation use case. Next, one should brainstorm data sources that might inform such inferences. Then, with data sources and measures selected, one should consider any additional inferences that need support. Then, one should generate a Validity Inference X Data Sources Matrix to ensure all inferences will be supported by some form of data. Next, one should collect the identified data whether it be qualitative or quantitative. Finally, one should examine the collected evidence in favor (or not) of the simulator's fidelity relative to the real-world task being simulated.

Table 1. Fidelity Evaluation Method.

1. Identify and define relevant validity inferences given number of task domains
2. Determine which measures would support the identified inferences
3. Identify and define relevant validity inferences given number of task domains and measure choice
4. Ensure support for all validity inferences
5. Generate Validity Inference X Data Sources Matrix to ensure all inferences will be supported by some form of data
6. Collect data
    a. Cognitive task analysis and/or fidelity requirements analysis
    b. Qualitative data about how prior research has defined and attempted high fidelity simulation of the task (i.e., literature review)
    c. Qualitative and quantitative information collecting perceptions, user feedback, and support for face validity (i.e., simulator cue rating worksheet; missing cue analysis)
    d. Quantitative information collecting behavioral and performance data (i.e., performance data)
7. Draw conclusions

This method supplements previous methods of fidelity suitability evaluation. Prior methods of quickly determining simulator utility have addressed considerations for simulator purchasers such as purchase cost, scalability, and reputation of manufacturer (Rooney et al., 2018). Other methods address the functionality of simulators, listing whether the functionality exists or not (Ruisi, 2017). Such methods have considerable value in highlighting logistical concerns and the existence of various functionalities. However, the tools advocated by Rooney and colleagues (2018) and Ruisi (2017) can be supplemented with a richer conceptualization of the veridicality with which training devices represent target procedures. That is, prior methods do prioritize measurement of the realism (fidelity) of procedure representations. Further, such methods did not involve the use of cognitive task analysis to identify the cues that need simulating. Typically, checklist tools to support procurement rely on subjective judgments of a simulator's support for various procedures. Not only does such a focus neglect the quality of procedure representation, it may also rely on subjective judgments about functionality. Such judgments often reflect psychological biases (e.g., confirmation bias, availability heuristics, etc.). The following paragraphs present a supplemental approach that integrates fidelity as a driving influence. Fidelity is a valuable differentiator for simulators that we believe can be leveraged to enable better decision-making. The proposed approach should supplement existing methods of functionality determination (e.g., Rooney et al., 2018; Ruisi, 2017), incorporating physical and functional fidelity with respect to both the existence and salience of physical cues, as derived through cognitive task analysis.

To test the proposed method, the contractor team first identified the relevant validity inferences (Figure 1, Figure 2). Validity inferences are a novel contribution to the fidelity literature from the efforts of TO 19. The contractor team wrote and plan to publish a manuscript detailing their conceptualization of fidelity evaluation as testing a set of validity inferences. In short, actual task performance (the jagged-edged ovals; Figure 1, Figure 2) is operationally defined through various forms of analyses (e.g., job, task) with respect to the involved cues, behaviors, and outcomes. Such performance-related factors constitute what is called the performance domain (i.e., the round ovals; Figure 1, Figure 2). Said cues, behaviors, and outcomes are all interdependent and inform each other as one completes the task. Observed measures (i.e., rectangles; Figure 2) attempt to measure the identified factors in the performance domain. Further, observed measures may be related to each other and even latent performance domain constructs. Finally, the latent performance domain constructs may be related to each other. The contractor team believed the relationship between the performance domain of a simulator and the performance domain of the real-world task (i.e., that which is being simulated) is the best approximation of observable fidelity (also known as "pragmatic fidelity"; Roza, 2004). Also, the contractor team distinguished between a proximal and distal simulator. Often, the real-world task is not measurable, necessitating the usage of two simulators: (1) a proximal simulator and (2) a distal simulator (or 'simulator simulator,' Meyer et al., 2012). The contractor team defined them as such due to the respective simulators' expected 'distance' from the real-world task along various metrics of fidelity. That is, a distal simulator's fidelity relative to the real-world task may be estimated transitively through its similarity to a more proximal simulator. For example, for various medical tasks, a proximal simulator might be a human cadaver and a distal simulator might be a 3D printed model (e.g., Huang et al., 2021). The fidelity of the 3D printed model may be estimated through the support of the inferences tested between it and the cadaver task. Overall, this framework of validity inferences helped identify several critical assumptions that are typically left unaddressed in fidelity research.



Note. Sim = Simulator

Figure 1. Identification of Validity Inferences Given the Number of Task Domains.

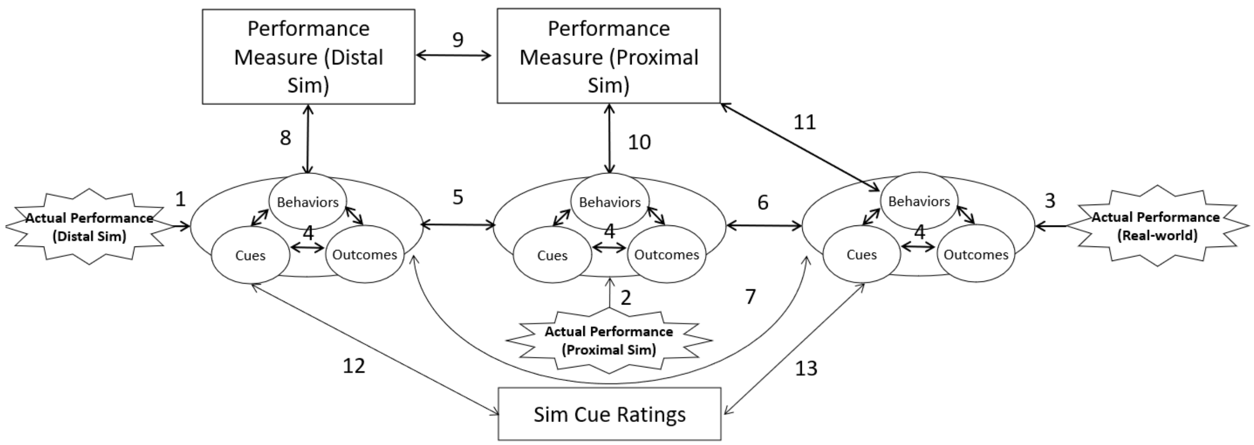
We defined the previous elements within our context of cricothyroidotomy simulator fidelity evaluation. In our context, the actual performance (real-world) is the actual cricothyroidotomy task being performed in the field on live patients. From that, the performance domain for actual performance (real-world) represented the identifiable cues, behaviors, and outcomes characteristic of successful cricothyroidotomy performance *on a live patient* (i.e., a list of task components generally collected from subject matter experts). Actual performance (proximal sim) is the process of one performing cricothyroidotomy on a simulator closely resembling the real-world referent; in our context, this was the process of one actually performing cricothyroidotomy on a porcine trachea simulator. The performance domain for the proximal simulator represented the cues, behaviors, and outcomes characteristic of successful cricothyroidotomy performance *on the proximal (i.e., porcine) simulator*. Finally, actual performance (distal sim) is the process of one performing cricothyroidotomy on a simulator that does not resemble the real-world referent as well as the proximal simulator is assumed to; that is, performing cricothyroidotomy on non-porcine simulators. The performance domain for the distal simulator represented the cues, behaviors, and outcomes characteristic of successful cricothyroidotomy performance *on the distal (i.e., non-porcine) simulator*.

Before selecting measures, one should identify the relevant inferences between the latent performance domains involved. For live patient surgical cricothyroidotomy, a task that makes data collection difficult, there is a need to identify the performance domains of a proximal simulator, distal simulator, and the real-world task (i.e., live patient cricothyroidotomy). Further, there is a need to support both the inference that the distal simulator's performance domain is related to the proximal simulator's performance domain (i.e., Inference 5) and that the proximal simulator's performance domain is related to the real-world task performance domain (i.e., Inference 6). Finally, there is the key inference being tested in this fidelity evaluation procedure: the extent to which the distal simulator performance domain is related to the real-world task performance domain (i.e., Inference 7).

Next, the contractor team identified the various methods to support such inferences. Inference 5 was testable using experimental protocols with in-person data collection whereas Inferences 6 and 7 were not. Because Inferences 6 and 7 involve the real-world task performance domain, which for the use case (live patient cricothyroidotomy) was impossible for behavioral data collection, the team decided to pursue questionnaire data. That is, creation of a simulator cue rating worksheet that asks respondents the extent to which cues in the distal simulator are realistic of the real-world task allows direct empirical support of Inference 7. There were plans for collection of similar data to support Inference 6, asking respondents to rate the realism of the proximal simulator relative to the real-world task, but the plans for such data collection were halted due to corona virus disease (COVID)-related restrictions on data collection. However, Inference 6 could still be supported by the conduction of a literature review that might reveal prior research's determination as to the similarity of the proximal simulator's performance domain with the real-world task performance domain (i.e., whether porcine simulated cricothyroidotomy is related to live patient cricothyroidotomy).

Methods of inference support attempted to synthesize prior fidelity evaluation methods in the team's choice of data sources. First, researchers followed recommendations from Holt and colleagues' (2012) report that identified the need to document the training capabilities of a system and its deficiencies. For testing the training capabilities, methodological recommendations were invoked from Bell and Waag (1998). Bell and Waag's (1998) approach to training system evaluation was to first consult with subject matter experts (SME) and ask them to provide self-reported perceptions, (2) determine if simulator training improved performance within the simulation environment, (3) determine if the learning that occurred in the simulation environment transfers to alternative simulation environments, and (4) determine if the learning that occurred transfers to the operational environment. Given how data collection in the operational environment of live patient cricothyroidotomy was impossible, work focused on replicating steps 1-3. Finally, the method returned to borrow from Holt and colleagues' (2012) report through the execution of a simulator cue rating questionnaire and a simulator cue deficiency report. All the data sources would provide support for the principal inference of interest: Inference 7. In sum, the choice in data sources resulted in a cognitive task analysis, experimental performance data, simulator cue rating data, literature review data, and missing cue analysis data. Thus, choice of measures was informed by but also expanded upon prior models through the identification of validity inferences and their need for support.

With the choice of measurement decided, there was a need to model the additional inferences. Given the decision to use behavioral performance data from the distal and proximal simulator performance domains, one must be mindful of additional inferences. For example, that both the performance measure for the distal (i.e., Inference 8) and proximal (i.e., Inference 10) simulators be representative of their respective performance domains. Further, to defend any claim that performance on the distal simulator is predictive of performance on the real-world task, there is a need to consider both the inference that it is related to the proximal simulator's performance measures (Inference 9) and that the proximal simulator's performance measures are representative of the performance domain of the real-world task (Inference 11). In essence, it must be ensured that completion time on a non-porcine simulator is significantly related to completion time on the porcine simulator and that completion time on the porcine simulator is representative of performing live patient cricothyroidotomy. Further, the inclusion of the simulator cue rating worksheet data assumes that the measure adequately samples the necessary cues for the real-world task (Inference 13) and assesses the perceived availability and realism of such cues in the distal simulator (Inference 12).



Note. Sim = Simulator

Figure 2. Identification of Validity Inferences Given the Number of Task Domains & Measure Choice.

We defined the previously described measures within our context of cricothyroidotomy simulator fidelity evaluation. Performance measure (distal sim) represented any (of multiple) performance measures that may be used when collecting data on the non-porcine simulator; in our context, this meant either a measure of completion time, commission errors, and omission errors on non-porcine cricothyroidotomy. Performance measure (proximal sim) represented the same measures, only captured while one is performing cricothyroidotomy on the proximal (i.e., porcine) simulator. The simulator cue ratings was a single self-report measure that assessed trainees' perceptions of the existence and realism of live patient cricothyroidotomy cues in the distal (i.e., non-porcine) simulator.

Next, the contractor team ensured all validity inferences would be supported with the proposed data collection method. It should be mentioned that methods of supporting inferences are distinct from methods of supporting hypotheses. Validity inferences are simply logical 'connective tissue' that highlight the various assumptions that should be monitored and tested using various data sources. When an inference is supported, it means there is data that assesses the validity of the claim. For example, Inference 7 may be supported as an inference but the data sources that support it may indicate no signs of fidelity. So, the contractor team defined all inferences and identified how each would be supported by one of the four proposed sources of data (Table 2). Next, the team used a Validity Inferences X Data Sources Matrix, mapping each inference to any number of data sources that would support it (Table 3). Together, such exercises ensured the existence of evidence supporting transitive logic that observed data assessing the distal simulator may in fact inform claims of the simulator's fidelity compared to the real-world task (Inference 7). Without such considerations, methods of evaluating fidelity may accidently make assumptions about the nature and relationships of observed data that can undermine the validity of results.

Table 2. Description and Method of Support for Validity Inferences.

Validity Inference	Description	Method of Support
1	The distal simulator performance domain is representative of performing cricothyroidotomy on a distal (i.e., non-porcine) simulator.	Supported by cognitive task analysis
2	The proximal simulator performance domain is representative of performing cricothyroidotomy on a proximal (i.e., porcine) simulator.	Supported by cognitive task analysis
3	The real-world task performance domain is representative of performing cricothyroidotomy on a live patient.	Supported by cognitive task analysis
4	Cues, behaviors, and outcomes associated with successful task performance are all interrelated.	Supported by cognitive task analysis
5	The extent to which cricothyroidotomy using the distal (i.e., non-porcine) and proximal (i.e., porcine) simulators are related.	Supported by a significant relationship between performance measures from the distal and proximal simulator
6	The extent to which cricothyroidotomy using the proximal (i.e., porcine) simulator and live patient are related.	Supported by prior research that might state porcine is a close approximation of live human patient (i.e., literature review)
7	The extent to which cricothyroidotomy using the distal (i.e., non-porcine) simulator and live patient are related.	Supported by estimations of physical (simulator cue ratings, missing cue analysis), behavioral (performance measures; assuming support of diagonal inference from cognitive task analysis), and outcome (cognitive task analysis) fidelity
8	The chosen measure of performance for the distal (i.e., non-porcine) simulator adequately represents the behaviors necessary for performance.	Supported by cognitive task analysis
9	The observed performance measures between the distal and proximal simulators are significantly related.	Supported by a significant relationship between completion times using either simulator
10	The chosen measure of performance for the proximal (i.e., porcine) simulator adequately represents the behaviors necessary for performance.	Supported by cognitive task analysis
11	The chosen measure of performance for the proximal (i.e., porcine) simulator is representative of live patient cricothyroidotomy's performance domain.	Supported by cognitive task analysis
12	The simulator cue ratings measure assesses the perceived availability and realism of cues in the distal (i.e., non-porcine) simulator.	Supported by participants' responses
13	The cues listed on the simulator cue ratings measure are the necessary cues for live patient cricothyroidotomy.	Supported by defining the list of cues present in the measure by using the list of cues present in the real task

Table 3. Validity Inferences X Data Sources Matrix.

Validity Inference	Cognitive Task Analysis	Experimental Performance Data	Simulator Cue Rating Worksheet	Literature Review	Missing Cue Analysis
1	X				
2	X				
3	X				
4	X				
5	X				
6		X			
7	X				
8		X			
9	X			X	
10				X	
11			X		
12			X		
13	X	X	X	X	X

The contractors collected the necessary data within a 3-phase approach. In Phase 1, the team conducted the cognitive task analysis (Appendix A). In Phase 2, the team conducted their training experiment, collecting the necessary performance data and simulator cue ratings data. In Phase 3, the team supplemented the simulator cue ratings data by collecting such ratings from trainers and procurement officers in the DoD; also, the team conducted the literature review (Appendix B) and missing cue analysis (Appendix C).

Additional data collection was anticipated to further support inferences of fidelity but was not completed due to COVID-related complications and low response rates. As previously mentioned, an element of Bell and Waag's (1998) training evaluation model discussed the importance of transfer of training to alternative simulation environments, not simply accumulation of skills from a pre- to post-test. To replicate this suggestion, the team designed an experimental protocol that would align with Bell and Waag's (1998) recommendations. Such data would further inform inferences 8-10. Due to COVID-related restrictions on in-person data collection, the execution of such a study was rendered impossible given increased precautions being taken at WPAFB, where the team planned to collect data. Also, the team identified a limitation of the simulator cue rating worksheet being its task-specificity, such that it could not be immediately applied to alternative real-world tasks of interest. As such, the contractor team developed a questionnaire to gauge fidelity perceptions in a distal simulator that would be a generalizable tool beyond cricothyroidotomy use cases. Unfortunately, when data was collected to validate the measure in Phase 3, the team received an insufficient number of responses to properly validate the scale ($N = 17$). The measure is still available for additional data collection and validation but was not used in subsequent fidelity evaluation analyses.

3.1 Phase 1 Method

3.1.1 Phase 1 Participants, Materials, Procedure

Information sources for task analytic investigations usually include formally documented task training materials, observational data from learning sessions, descriptive interviews with subject matter experts that represent one of several possible approaches, including unstructured interviews, concept maps, conceptual graph analysis, critical decision analysis, constructed scenarios, repertory grids, and/or knowledge audits as well as observation methods such as process tracing, direct observation-questioning, and expert/novice comparisons (Klein & Militello, 2001). Other sources of data include criteria for evaluation present within performance appraisal systems. The optimal number of respondents for such a method is highly dependent on the availability of the sample. Sample size is a complex consideration. Militello and Hutton (1998) stress that training requirements, data coding and transformation efforts, and availability of SMEs often result in small sample sizes for consultation. For our investigation, we interviewed nine medical trauma training experts, acquiring detailed task information about surgical cricothyroidotomy and validating time estimates and potential errors associated with each step.

To guide task selection, the MARCH-PAWS mnemonic indicates the critical nature of airway management during tactical combat casualty care. Tasks included nasopharyngeal airway insertion and surgical cricothyroidotomy (using the Cric-Key evaluation and inserting an endotracheal tube for airway establishment). The Center for Army Lessons Learned Handbook (CALL, 2017) provided an initial listing of the major tasks and subtasks for the procedure.

The contractor team contacted experts in surgical cricothyroidotomy instruction. The purpose of this outreach was to supplement the task analysis created from existing scientific literature journal articles and reports. The contractor team led contributors through the task analysis step-by-step, making notes about the changes that were recommended. Specific issues included the sequence of steps indicated, the importance of each step, the expected time required for each step, errors that were possible during execution of each procedural step, and criteria for procedure success. The contractor team merged the information gained from the literature with the information gained through the consultation with SMEs. The team then sent out the revision to the participants for their concurrence and approval.

3.2 Phase 1 Results

The task analysis refined the list and provided a detailed and relevant task description to use for comparative evaluation across CEM platforms. By the conclusion of the first phase of research, we had completed our initial analysis and prepared several deliverables:

- As an outgrowth of the task analytic process, we constructed a hierarchical and functional listing of component skills associated with the surgical cricothyroidotomy procedure (See Appendix A).

- To supplement the task analysis, we provided an error analysis framework, listing potential errors of omission and commission. This underscored the relative importance of each task.
- From the prior deliverables, the ultimate product was a fidelity matrix (Simulator Cue Rating Worksheet) that allowed research participants to evaluate the selected CEMs for their potential to provide sensory cues relevant to all surgical cricothyroidotomy tasks.

Completion of the task analysis informed measure development; completion of an error analysis (listing of potential errors associated with each task step as well as documentation of errors committed by participants during testing); documentation of conventional performance metrics such as procedure completion time and task accuracy; incorporation of eye tracking measurement that included fixation and scan path efficiency analyses; investigation of existing fidelity assessment tools and development of novel questionnaires for fidelity measurement; and appropriation and refinement of self-efficacy measurement instruments.

Following the refinement of the task analysis, we integrated the task cue, readiness, and error information into the Simulator Cue Rating Worksheet to be used during data collection in the human performance experiment with experienced and inexperienced participants.

3.3 Phase 2 Method

3.3.1 Phase 2 Data Collection to Supplement Fidelity Ratings

Following the development of the cognitive task analysis and creation of the Simulator Cue Rating Worksheet during Phase 1, we supplemented the task analytic data with performance data as medical SMEs familiar with combat casualty care performed surgical cricothyroidotomy on both selected CEM trainers (Laerdal ALS and SynDaver Adult Cricothyroidotomy Trainer). The importance of including performance measures is underscored by other military simulator fidelity evaluations, including the seminal treatments of fidelity by Baum et al. (1982), Bell and Waag (1998), and Holt et al. (2012).

Selection of SMEs to serve as task performers and evaluators was undertaken to inform a subsequent procedure of cue consensus. We initially identified emergency medical technicians who had knowledge and recent practical experience to candidly evaluate the capabilities of each simulator and who had the applied skill to attempt the target procedure on each.

During Phase 2, we gathered performance data that supplemented our analytic comparison of CEMs in Phase 1. Consequently, our expected outcomes of Phase 2 were:

- Speed and accuracy performance data from inexperienced and experienced participants
- Self-report fidelity perceptions data
- Statistical comparisons of collected metrics to better highlight differences in simulator fidelity

3.3.2 Phase 2 Data Collection Experimental Design

The human performance experimental research was separated into two data collection sessions: one for task experts and one for task novices. In each case, the experimental design was the same: variables were manipulated and measured in accordance with a 7 x 2 mixed experimental design, where training simulator used was manipulated as a between-groups variable (participants were randomly assigned to the Laerdal ALS Manikin or the SynDaver Adult Cricothyroidotomy Trainer). Task performance session constituted a within-group variable, with participants performing simulated cricothyroidotomy seven times. Given prior research, we tested the following hypotheses:

- Hypothesis 1: ***Procedure completion time*** will decrease linearly over time
- Hypothesis 2: ***Number of commission errors*** will decrease from pretest to posttest (Proctor, 2005)
- Hypothesis 3: ***Number of omission errors*** will decrease from pretest to posttest (Proctor, 2005)
- Hypothesis 4: ***Knowledge test scores*** will increase following training (Kahol, Vankipuram, & Smith, 2009)
- Hypothesis 5: ***Rated self-efficacy*** will be negatively associated with task knowledge (Mahmood, 2016; Moore & Healy, 2008)
- Hypothesis 6: ***Rated self-efficacy*** will be more variable for individuals with lower levels of demonstrated competence (i.e., higher completion time) (Gignac & Zajenkowski, 2020)
- Hypothesis 7: ***Tactile cues*** will be rated as most realistic for the SynDaver trainer (Mund, 2011)
- Hypothesis 8: ***Visual cues*** will be rated most realistic for the Laerdal ALS trainer (Johnson, 2017)
- Hypothesis 9: ***Auditory cues*** will be rated as less realistic than tactile and visual cues.
- Hypothesis 10: ***Frequency of recorded eye fixations*** will be greatest for the tracheal region (Hermens, Flin, & Ahmed, 2013)
- Hypothesis 11: ***Total recorded eye fixations*** will decrease from pretest to posttest

3.3.3 Phase 2 Participants

For the first experimental phase, participants with experience conducting surgical cricothyroidotomy were trained and tested. Experts were defined as first responders drawn from Dayton, OH area fire departments who had performed surgical cricothyroidotomy on a simulator or an actual person within the previous twelve months. For the second experimental phase, inexperienced participants were defined as students from Dayton area medical training programs at Wright State University and the University of Dayton who had completed at least one class in anatomy or physiology, but who had not performed surgical cricothyroidotomy at any time in the previous twelve months. Following from a power analysis (alpha = .05, medium effect size, beta = .80), the target sample size for experts was 10; the target sample size for novices was 40. The lower number for experts reflects the expected homogeneity inherent within that sample. Ultimately, 14 experienced participants and 38 inexperienced participants were recruited and tested.

Each participant volunteered individually for a 90-minute time slot. The experimenter confirmed that the participant was a United States citizen (necessary for military base access) and that he or she did not require eyeglasses to perform a near-work task. He then emailed them to confirm their time and to provide a map showing a meeting place off base. On the day of participation, the experimenter met each participant and escorted him or her onto the base. Following completion of their participation, the experimenter transported each participant back to his or her car. Participants received no compensation for their participation.

3.3.4 Phase 2 Materials

During the experiment, participants completed eight questionnaires. After reading the informed consent form, participants completed a demographics and experience questionnaire. The demographics and experience questionnaire included items such as participant age, number of anatomy and physiology classes completed, number of surgical cricothyroidotomy procedures completed during the previous year, and extent of experience as a first responder (if appropriate). It also enabled participants to indicate any medical certifications they held.

Participants then completed a ten-item multiple-choice knowledge pretest to determine their mastery of anatomical, physiological, and procedural elements of surgical cricothyroidotomy. The items on the quiz were constructed according to the results of a cognitive task analysis and were approved for inclusion by a subject matter expert with over 30 years' experience teaching and performing surgical cricothyroidotomy. A parallel forms version of the test was completed by each participant following their task training sessions.

After the parallel forms version of the cricothyroidotomy knowledge test, participants completed a self-efficacy rating scale that was constructed according to recommendations by Bandura (2006). The form required participants to estimate their level of confidence (from 0 = "cannot do at all" to 100 = "highly certain can do,") in performing each of the 22 steps of the surgical cricothyroidotomy procedure taken from the cognitive task analysis.

Following the final performance test, participants completed a Simulator Cue Rating Worksheet. As a central element of the fidelity rating procedure, the scale required participants to estimate the realism of each of 44 auditory, visual, and tactile sensory cues considered necessary to fully complete the surgical cricothyroidotomy procedure. As with the elements in the other questionnaires, the cues were validated by a group of nine subject matter experts who had extensive knowledge and practical experience performing and teaching surgical cricothyroidotomy. Rating inputs were structured as Likert ratings from 0 = "Cue is not available from the simulator" to 4 = "Simulator is extremely realistic for detection of this cue." The Simulator Cue Rating Worksheet also allowed participants to provide open-ended comments justifying their numeric ratings if desired.

Next, participants completed the Situation Awareness Rating Technique (SART) questionnaire; (Selcon & Taylor, 1990). That questionnaire required them to estimate training situation instability, complexity, and variability; arousal; attention concentration and division; spare mental capacity; information quantity; and situation familiarity. After completing the SART, participants completed the National Aeronautics and Space Administration - Task Load Index (NASA-TLX) (Hart, & Staveland, 1988) to estimate their level of perceived cognitive workload across six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration).

The final questionnaire was designed to elicit participants' opinions about the experiment. They estimated how helpful the simulator was for learning the procedure (from "not helpful at all; I feel no better prepared to perform the procedure" to "extremely helpful; I gained considerable expertise performing on it"). The questionnaire also enabled participants to rate general helpfulness (reverse coded, from 1 = extremely helpful to 5 = not helpful at all) of the simulator they used for visual, tactile, auditory, structural, and functional task cues. There were also open-ended questions pertaining to differences between the porcine trachea and the simulator, estimates of predicted difficulty between the simulator and a hypothetical live patient, and space for participants to provide simulator improvement suggestions.

In addition to the questionnaires, participation required that individuals perform the surgical cricothyroidotomy task twice on a porcine trachea (pre- and post-test) as well as on the assigned training simulator. Porcine tracheas were purchased from a local meat processing facility in compliance with Federal Department of Agriculture (FDA) regulations. They were frozen until ready for use, then thawed and placed on a fabricated task board. The tracheas were covered with 6" x 6" squares of chamois sheepskin to simulate outer skin. After use, porcine tracheas and related materials were collected in infectious waste bags and delivered to the base officer for appropriate disposal.

Laerdal SimMan ALS LiveShock ™ Simulator – Half of the participants in each experimental group were randomly assigned to be trained using a Laerdal SimMan ALS Liveshock manikin simulator (Model No. 235-03350). The simulator allows individuals to practice surgical cricothyroidotomy by making incisions through replaceable silicone neck skin and through a tape layer that represents the cricothyroid membrane. Anatomical structures such as the hyoid bone are represented underneath the replaceable neck skin, including a tracheal opening. For this experiment, the manikin was not connected to a computer.

SynDaver Adult Cric Trainer ™ – The other half of the participants were randomly assigned to training using the SynDaver Labs SynAtomy Adult Cric Trainer (Model No. 160440). The trainer consists of a plastic base, muscular neck form, skin overlay, cric carriage, silicone O-ring, cricothyroid cartilage, and replacement tissues. The trainer has anatomical features such as an oral and nasal passages, hyoid bone, thyroid cartilage, cricoid cartilage and cricoid membrane.

Cric-Key kit description – The Control-Cric™ (Cric-Key) tool kit (Product No. 900783) included a Cric-Knife scalpel with attached tracheal hook, Melker airway introducer with inserted hard plastic bougie and attached 5.5mm endotracheal line and cuff, rubber strap and wedge for stabilizing the airway following insertion (not used for this procedure). Participants also used a bag/valve/mask device for ventilating the airway at the end of the procedure. The device was not included as a part of the Control-Cric™ kit. The Control-Cric system has been recommended by TCCC for training and performing surgical cricothyroidotomy.

Participants wore standard nitrile latex gloves and protective masks while completing the procedure. During the pretest, training and posttest sessions, participants also wore Tobii Glasses 2 eye tracking glasses to record their visual fixations and scan patterns.

3.3.5 Phase 2 Procedure

Upon arrival at the laboratory, participants read through the informed consent form and completed the background questionnaire and the cricothyroidotomy knowledge pre-test. All participants began by providing written informed consent. They were also asked to complete a background demographics questionnaire specifying their level of experience with MARCH-PAWS procedures. They then donned the Tobii Glasses 2 eye tracking glasses and underwent gaze calibration as specified within the Tobii manual. After calibration, the experimenter directed the participant to the porcine trachea task station. Then, the participant attempted to complete the pre-test cricothyroidotomy procedure on the porcine trachea.

During each session, an SME was asked to perform surgical cricothyroidotomy seven times. After performing the pretest procedure using an animal model, participants were trained on the procedure didactically and practically, performing the same procedure five times on one of the two simulators (randomly determined). The final procedure was a posttest, performed without guidance on a porcine model. Performance measures included speed of procedure execution, errors committed during the procedure (including errors of commission as well as those of omission), and eye tracking measures of fixation and scan path efficiency. In cases where the simulator did not support completion of a task, the failure was documented as well. Importantly, measures of cognitive workload and Levels 1, 2, and 3 situation awareness were incorporated into the data collection procedure. During this process, we relied substantially on the completed task analysis to determine error categories.

After the pre-test, participants were guided through standardized training by the subject matter expert. Training consisted of explanation of anatomical features, criteria for procedure initiation, description of procedural steps, and possible complications. The subject matter expert referred to wall-mounted illustrations during instruction. During training, the expert described the correct procedural steps and sequences to correctly perform the cricothyroidotomy procedure based on the TCCC task skill sheet. The task began with “can’t intubate, can’t ventilate” assumed and Control-Cric tools available within a nylon bag. Participants were told to emphasize task speed and accuracy equally. The task steps included the following:

Step 1. Test equipment for functionality

Step 2. Feel for anatomical landmarks with non-dominant hand (perform the “laryngeal handshake”)

Step 3. Trace downward on trachea with finger, feeling for the step off location of cric membrane

Step 4. Make .5 inch vertical incision through outer skin

Step 5. Spread subcutaneous tissues with tracheal hook or rotate blade (hook use optional)

Step 6. Make .5 inch horizontal incision through cricothyroid membrane

Step 7. Confirm entry into trachea with finger

Step 8. Insert Cric-Key Melker introducer with bougie inserted

Step 9. Remove bougie from introducer

Step 10. Inflate attached endotracheal cuff

Step 11. Attach BVM bag valve and ventilate; watch/listen for bilateral lung rise and air movement

During training, participants performed the cricothyroidotomy procedure using their assigned simulator five times. The instructor guided them during each performance, correcting actions that were inaccurate or inappropriate. Simulator components were replaced (i.e., simulated skin and cricothyroid membrane materials were replaced) after each training session. Following the fifth simulator training session, participants completed the cricothyroidotomy knowledge post-test and the self-efficacy rating scale. Next, participants returned to the porcine trachea station and performed the cricothyroidotomy procedure again as a post test. After completing the post test, participants were directed to complete the remaining self-report measures (i.e., Simulator Cue Rating Worksheet, SART, NASA TLX, and opinion questionnaire).

After completing the last questionnaire, participants were debriefed concerning the goals of the experiment, expected findings, and expected benefits of the research for warfighters and medical personnel. All participants were fully debriefed in adherence to AFRL Institutional Review Board (IRB) approval requirements. They were then escorted back to their vehicles, thanked, and dismissed.

3.4 Phase 2 Results

Following data collection, experimenters first electronically coded all questionnaire data. They then transcribed all training and test performance videos, noting procedural completion time in seconds, errors of omission and errors of commission (including actions that violated optimal sequence, duration, or magnitude). Last, they used Tobii Pro Lab software to analyze eye tracking data, noting fixations and scan patterns during training and test sessions. After all coding was complete, data were merged into a master Microsoft Excel spreadsheet that was formatted for repeated measures analyses. The spreadsheet was then ingested within the Statistical Package for the Social Sciences (Version 24) to complete analyses.

Analyses of eye tracking performances were accomplished by focusing on focal fixation patterns and scan patterns, as well as an eye scan path efficiency score that reflected rated adherence to an optimal scan pattern. Fixation data were prepared for analysis by manually examining each recorded fixation noted in the Tobii Pro Lab software using the "I-VT Attention" filter. That filter sets the velocity threshold parameter to 100 degrees/second, appropriate for constant or frequent head motion resulting from head-worn eye trackers such as the Tobii Glasses 2. Eye scan path efficiency data were prepared by manually reviewing and coding scan path videos within Tobii Pro Lab and comparing the coded scan paths to an optimal path. The rated paths were then subjected to a scoring algorithm that accounted for fixation sequencing and overall efficiency of eye fixations.

- Eye tracking analyses:
    - Total Eye Fixations – Tabulate eye fixations occurring during procedure
    - Area of First Fixation – Area where first fixation occurred after procedure start
    - Eye Fixation Proportion – Proportion of fixations for eleven regions of interest
    - Eye Path Sequence Efficiency – Score (0-7) reflecting optimality and efficiency of eye scan path

Addressing eye tracking hypotheses required experimenters to review performance videos of each participant as they performed the procedure. Participants in the experienced group performed the procedure without continuous guidance from the trainer; therefore, error tabulation was straightforward. In contrast, participants in the inexperienced group were coached by the trainer during the training performances. For this reason, experimenters coded errors for only the porcine pretest and posttest. Errors of commission were recorded according to recommendations by Swain and Guttman (1983). They included superfluous actions as well as actions that represented violations of duration, timing, and sequence. Experimenters reviewed the results of the cognitive task analysis to jointly develop a rubric whereby the errors could be identified and tabulated.

Experimenters verified completeness of the dataset by documenting missing data and outliers. Statistical outliers were defined as data points falling outside three standard deviations from the mean. All outlying data points were examined and judged to represent normal extremes of trainee performance; therefore, none were excluded from ensuing analyses. Further, we examined the normality of key study variables, checking the skew and kurtosis. All variables demonstrated adequate normality distributions (i.e., absolute value of skew did not exceed 2.0 and absolute value of kurtosis did not exceed 7.0). Following data cleaning and identification of missing and outlying data points, experimenters completed data analysis, adhering to a systematic plan that included *a priori* hypothesis testing and *post hoc* data trend identification. We implemented a combination of descriptive and inferential statistical techniques to analyze our findings. Where appropriate, we also employed distribution-free or nonparametric statistics to analyze questionnaire data that represented nominal or ordinal scaling. Group means representing performance for particular simulators were compared. We also relied on correlations to assess relations among task procedure performance measures and SME experience levels with combat casualty care and experience with the target procedure.

First, descriptive statistics were generated to determine measures of central tendency (mean for normally distributed variables; median for distribution-free variables) and variation (standard deviation and standard error). As a step in the statistical analyses, means were stratified by experimental condition and reported below as appropriate for inferential statistical testing.

3.4.1 Descriptive Statistics

The mean age in the inexperienced participant sample was 26.87 (standard deviation = 7.23), whereas the mean age in the expert sample was 32.21 (standard deviation = 10.30). The average number of anatomy/physiology classes completed by participants in the inexperienced sample was 4.42 (standard deviation = 2.98), whereas the average number of anatomy physiology classes completed by participants in the experienced sample was 2.57 (standard deviation = 1.83). In the inexperienced sample, no participants had been trained to perform surgical cricothyroidotomy, whereas in the experienced sample, six participants had performed cricothyroidotomy on actual patients and all others in the experienced group had performed the procedure on simulators or animals within the prior year.

We obtained the means and standard deviations for key study variables in the combined sample, the experts-only sample, and the novices-only sample (see Table 4).

Table 4. Sample Means and Standard Deviations for Experienced, Inexperienced, and Overall.

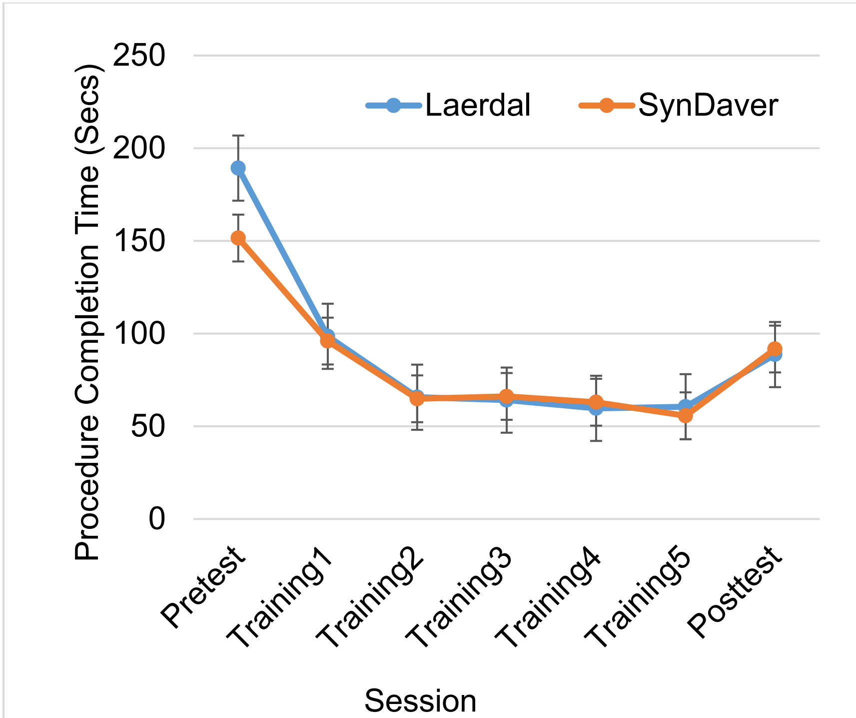
Study Variable	Experienced Mean (SD)	Inexperienced Mean (SD)	Overall Mean (SD)
Procedure Completion Time, Pretest	170.43 (59.18)	220.13 (74.68)	203.34 (75.19)
Procedure Completion Time, Session 1	97.29 (27.28)	173.18 (69.85)	149.40 (71.02)
Procedure Completion Time, Session 2	65.29 (15.78)	104.32 (32.44)	91.83 (34.87)
Procedure Completion Time, Session 3	65.14 (16.59)	78.58 (21.13)	73.70 (21.88)
Procedure Completion Time, Session 4	61.35 (18.49)	72.92 (20.50)	68.70 (21.28)
Procedure Completion Time, Session 5	58.14 (14.93)	64.79 (14.14)	62.02 (15.68)
Procedure Completion Time, Posttest	90.21 (34.33)	82.55 (21.42)	83.79 (25.85)
Omission Errors, Pretest	1.57 (.85)	4.84 (1.95)	3.86 (2.28)
Omission Errors, Session 1	2.00 (.88)		
Omission Errors, Session 2	2.00 (.88)		
Omission Errors, Session 3	2.07 (.92)		
Omission Errors, Session 4	2.21 (1.12)		
Omission Errors, Session 5	2.07 (1.14)		
Omission Errors, Posttest	1.00 (1.04)	.61 (.68)	.72 (.79)
Commission Errors, Pretest	1.86 (1.17)	3.34 (1.85)	2.89 (1.80)
Commission Errors, Session 1	1.50 (1.45)		
Commission Errors, Session 2	1.07 (.92)		
Commission Errors, Session 3	1.21 (.97)		
Commission Errors, Session 4	1.50 (.94)		
Commission Errors, Session 5	1.35 (.93)		
Commission Errors, Posttest	.71 (.73)	.71 (.80)	.71 (.76)
Knowledge Score, Pretest	6.57 (1.09)	6.13 (1.56)	6.16 (1.59)
Knowledge Score, Posttest	7.79 (1.42)	7.18 (1.06)	7.24 (1.42)

Eye Tracking, Total Fixations (Pretest)	200.42 (81.99)	333.71 (183.98)	286.46 (170.25)
Eye Tracking, Total Fixations (Posttest)	142.83 (41.85)	139.14 (64.36)	137.97 (58.58)
Eye Tracking, Scan Efficiency (Pretest)	4.91 (1.04)	2.18 (.87)	3.50 (1.71)
Eye Tracking, Scan Efficiency (Session 1)	4.73 (1.01)	4.70 (1.43)	5.00 (1.48)
Eye Tracking, Scan Efficiency (Session 2)	5.18 (.75)	5.04 (1.34)	5.27 (1.27)
Eye Tracking, Scan Efficiency (Session 3)	5.09 (.70)	5.32 (1.28)	6.00 (.77)
Eye Tracking, Scan Efficiency (Session 4)	4.91 (.94)	4.91 (1.18)	5.27 (.79)
Eye Tracking, Scan Efficiency (Session 5)	4.91 (.70)	4.98 (1.15)	5.45 (.69)
Eye Tracking, Scan Efficiency (Posttest)	5.09 (.94)	4.88 (1.24)	5.00 (1.00)

Note: Shaded cells indicate no error data analyzed for inexperienced participants' training sessions.

3.4.2 Task Performance Hypothesis Testing

Hypothesis 1 stated that procedure completion time would decrease linearly over time. To test this hypothesis, we computed a 7 (session) X 2 (training simulator) mixed factor analysis of variance with trend analysis for each expertise group (see Figure 3). For experienced participants, the analysis revealed no significant interaction or simulator main effect ($p > .05$); however, there was a main effect of experimental session, $F(2.42, 87.08) = 88.36$, $p < .001$, partial $\eta^2 = .71$, observed power = 1.00. Follow-up trend analyses showed that the main effect reflected linear, $F(1, 36) = 202.89$, $p < .001$, partial $\eta^2 = .85$, observed power = 1.00, and quadratic effects, $F(1,36) = 101.44$, $p < .001$, partial $\eta^2 = .74$, observed power = 1.00. In each case, the Mauchly test of sphericity was significant, so we applied the Greenhouse-Geisser correction. Thus, Hypothesis 1 was supported.

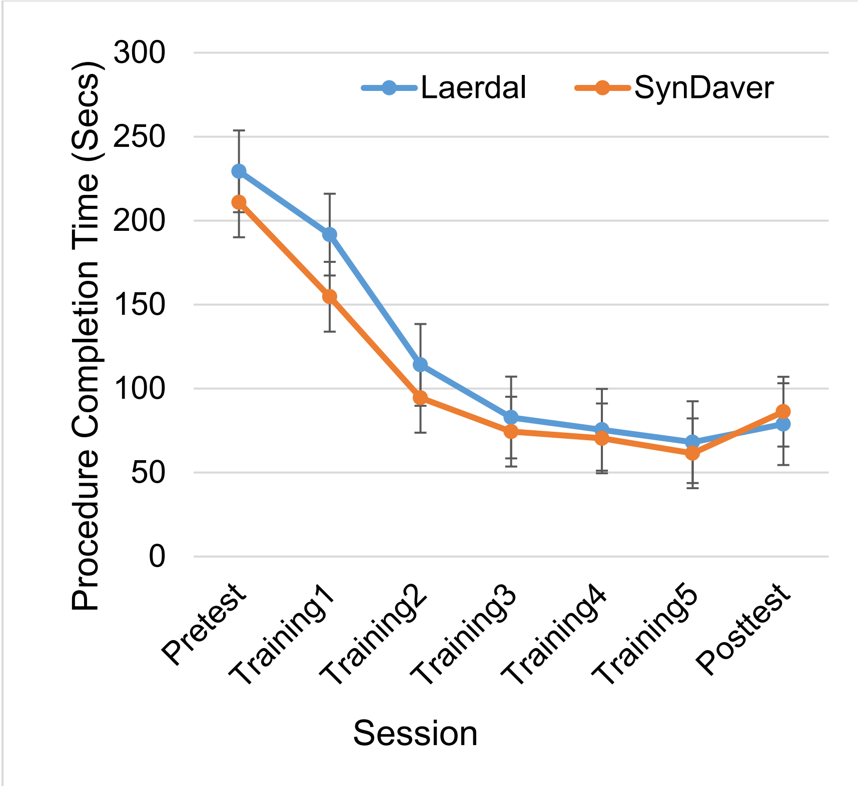


Note: Error bars represent standard error

Figure 3. Average Experienced Procedure Completion Time by Simulator and Session.

Similarly, our 7 X 2 mixed factor analysis for the inexperienced group showed no significant interaction of simulator main effect ($p > .05$). There was a main effect of experimental session, $F(1.67, 19.99) = 33.12$, $p < .001$, partial $\eta^2 = .73$, observed power = 1.00. Trend analyses again showed significant linear, $F(1, 12) = 55.94$, $p < .001$, partial $\eta^2 = .82$, observed power = 1.00, quadratic, $F(1, 12) = 45.46$, $p < .001$, partial $\eta^2 = .79$, observed power = 1.00, and cubic, $F(1, 12) = 5.21$, $p = .042$, partial $\eta^2 = .30$, observed power = .56, effects. As before, the Mauchly test was significant, necessitating the Greenhouse-Geisser correction (see Figure 4).

In addition to the 7 x 2 Analysis of Variances (ANOVA), we computed one-way ANOVA to directly compare pretest and posttest procedure completion times between experienced and inexperienced participants. Results indicated that average pretest times were significantly faster for experienced participants than for inexperienced participants, $F(2, 49) = 5.02$, $p = .01$. Posttest procedure completion times were not significantly different, $p > .05$.

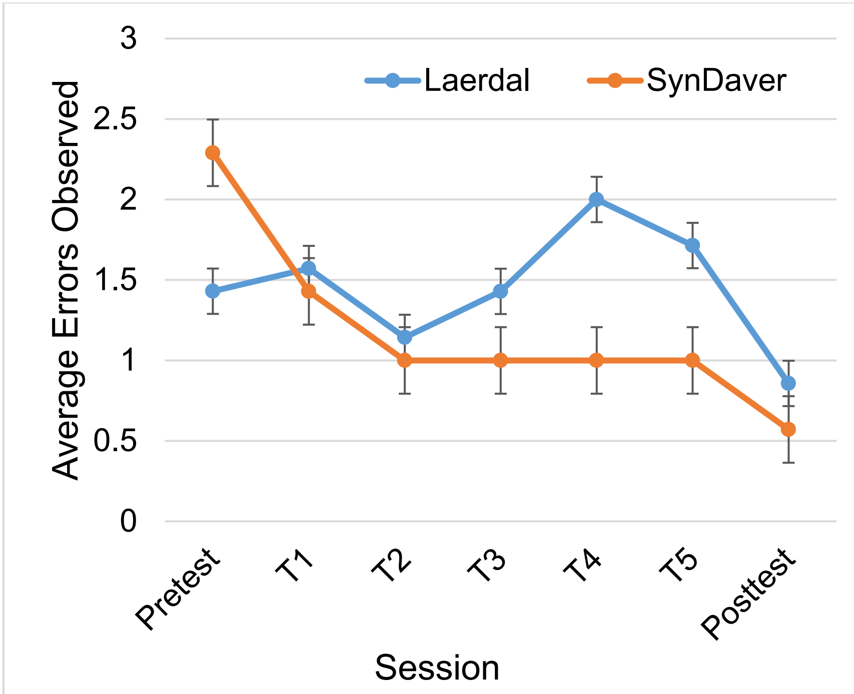


Note: Error bars represent standard error

Figure 4. Average Inexperienced Procedure Completion Time by Simulator and Session.

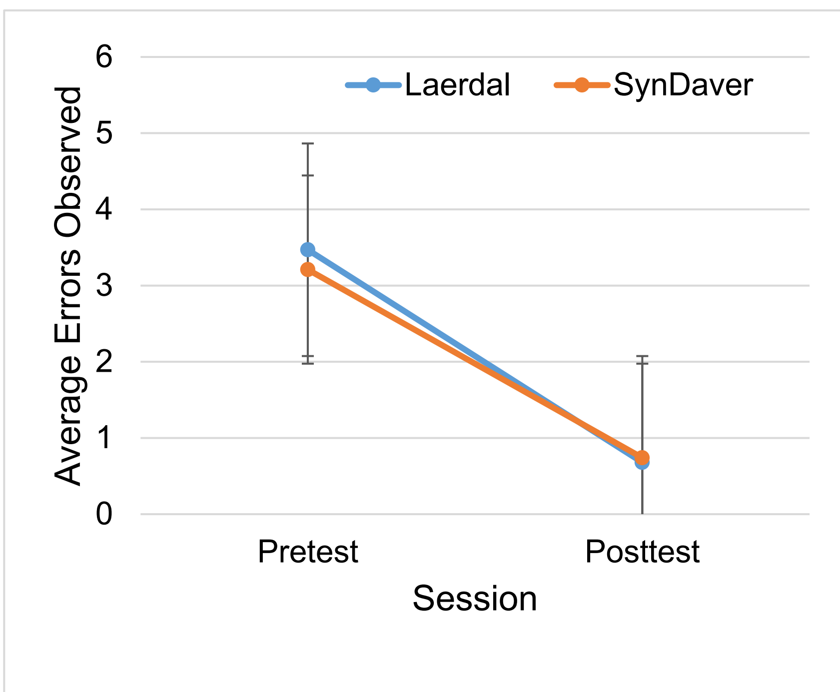
Hypothesis 2 stated that errors of commission would decline with increasing task experience. For experienced participants, we tested the hypothesis by computing a 7 (Performance Session) by 2 (Simulator) ANOVA with recorded commission errors as the dependent variable. Results indicated a non-significant interaction and non-significant main effects ($p > .05$); therefore, the hypothesis was not supported for the expert group. For inexperienced participants, we computed a 2 (Test Session) by X 2 (Simulator) ANOVA. That analysis showed no significant interaction or simulator main effect; however, there was a significant session main effect, $F(1, 36) = 81.74$, $p > .001$, partial $\eta^2 = .69$, observed power = 1.00. That effect showed a reduction of errors from pretest to posttest, supporting the hypothesis (see Figures 5 and 6).

A comparison of commission errors between experienced and inexperienced participants showed that the average pretest error rate for inexperienced participants was significantly higher than the rate for experienced participants ($p = .027$); however, posttest comparisons showed that the difference was eliminated with training ($p > .05$).



Note: Error bars represent standard error

Figure 5. Average Experienced Commission Errors by Simulator and Session.

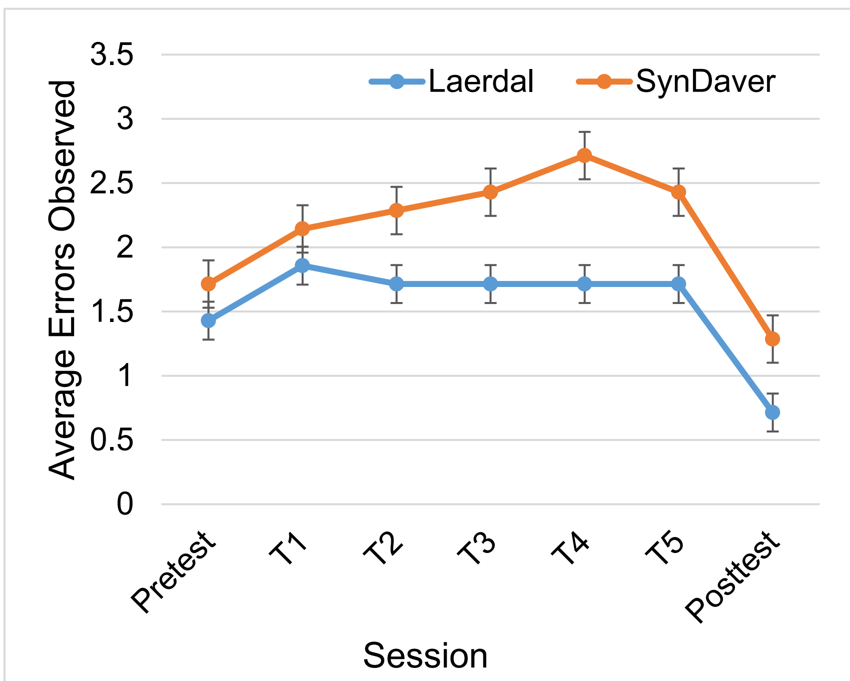


Note: Error bars represent standard error

Figure 6. Average Inexperienced Commission Errors by Simulator and Session.

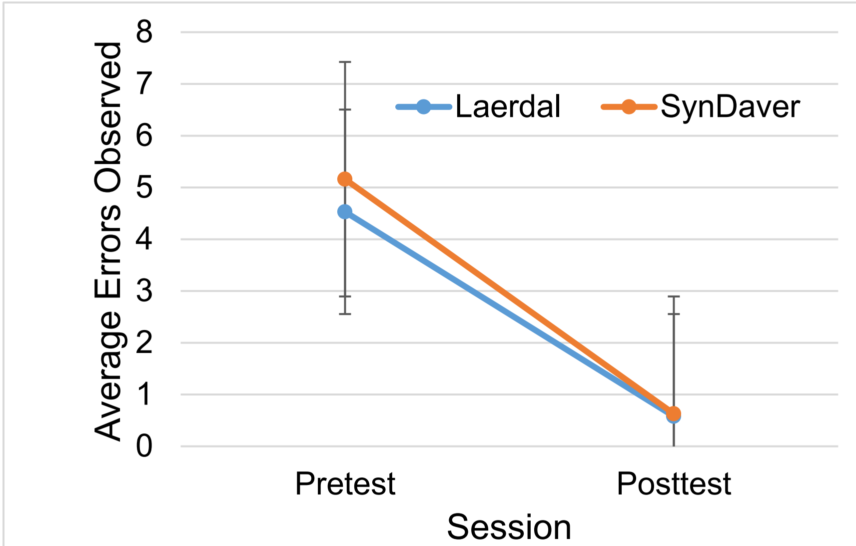
Hypothesis 3 stated that observed errors of omission would also decrease with increasing task experience. We tested this hypothesis separately for inexperienced and experienced groups, mirroring our approach for errors of commission. For experienced participants, we tested the hypothesis by computing a 7 (Performance Session) by 2 (Simulator) Analysis of Variance with recorded omission errors as the dependent variable. Results indicated a non-significant interaction and non-significant main effect of simulator ($p > .05$). The session main effect was significant, $F(3.09, 37.12) = 5.16$, $p = .004$, partial $\eta^2 = .30$, observed power =.90. Follow-up trend analyses showed that the pattern of means was not linear. Instead, errors generally decreased across sessions in a quadratic, $F(1,12) = 11.51$, $p = .005$, partial $\eta^2 = .49$, observed power = .88, and cubic, $F(1, 12) = 5.47$, $p = .037$, partial η^2=.31, observed power =.58, fashion.

For inexperienced participants, we computed a 2 (Test Session) by 2 (Simulator) ANOVA. That analysis showed no significant interaction or simulator main effect; however, there was a significant session main effect, $F(1, 36) = 148.21$, $p > .001$, partial $\eta^2 = .81$, observed power = 1.00. That effect showed a reduction of errors from pretest to posttest, supporting Hypothesis 3 (see Figures 7 and 8).



Note: Error bars represent standard error

Figure 7. Average Experienced Omission Errors by Simulator and Session.



Note: Error bars represent standard error

Figure 8. Average Inexperienced Omission Errors by Simulator and Session.

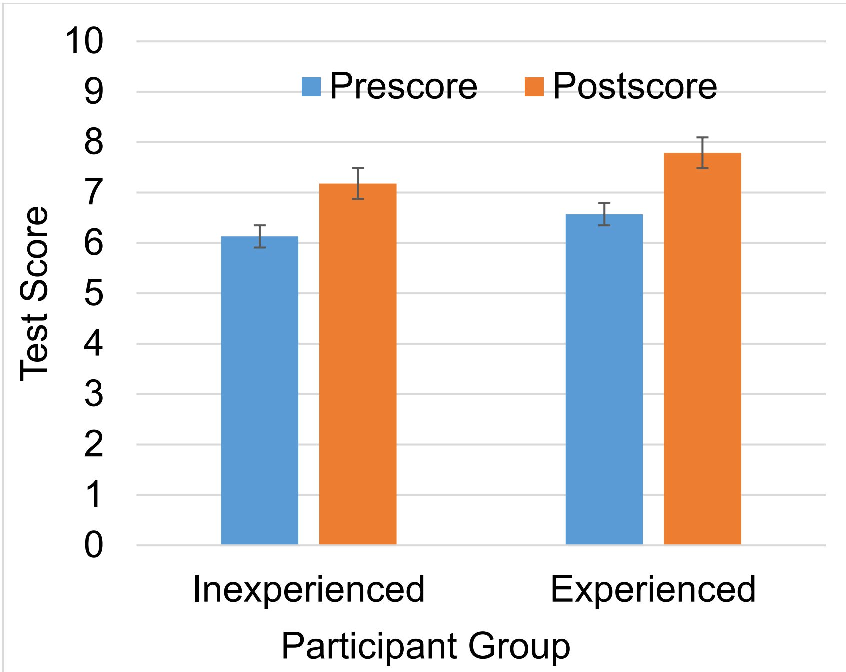
Comparing omission errors between experienced and inexperienced participants showed that the experienced participants' pretest average omission error rate was significantly lower than inexperienced participants' average ($p < .001$); however, the experienced posttest omission rate was significantly higher than inexperienced ($p = .005$)

The results of the task performance hypothesis testing demonstrated value as an empirical test of simulator fidelity and were published in the 2021 Proceedings of the Human Factors and Ergonomics Society Health Care Symposium (Bliss, Etherton, Hodge, & Winner, 2022a). Also, a subset of the analyses was presented as a poster at the Human Factors and Ergonomics Society annual conference (Bliss et al., 2021).

3.4.3 Confidence-Competence Hypotheses

The relationship between self-efficacy and performance has been studied in a variety of contexts. Researchers have interpreted competence by referring to cognitive test scores as well as traditional task performance measures. In the current research, we associated self-efficacy ratings with knowledge test scores as well as procedure completion time and errors.

Our first step was to determine whether task training led to increases in cognitive (knowledge) test scores. Hypothesis 4 stated that cognitive test scores would increase following training for both novice and expert samples. To test this hypothesis, we conducted *t*-tests between pretest and posttest scores for both samples. We observed a significant increase in cognitive test scores for both the inexperienced, $t(37) = -3.77$, $p = .001$) and experienced, $t(13) = 2.18$, $p = .048$, samples; see Figure 9). Thus, Hypothesis 4 was supported.



Note: Error bars represent standard error

Figure 9. Cognitive Test Scores Compared by Expertise Level and Administration Period.

Hypothesis 5 stated that self-efficacy would be negatively related to task knowledge. To test this hypothesis, we computed two correlations involving the pre- and post-training knowledge tests and self-efficacy score. The correlation between self-efficacy and pre—training task knowledge was negative and statistically significant ($r = -.30$, $p < .05$). The correlation between self-efficacy and post-training task knowledge was non-significant ($r = .01$, $p > .05$). Thus, participants with less task knowledge before training were more likely to have higher self-efficacy than those with more task knowledge before training. However, there was no relationship between post-training knowledge and self-efficacy. Thus, Hypothesis 5 was partially supported.

Further, we conducted multiple exploratory analyses to examine the effect of simulator or experience. So, we computed *t*-tests as well as correlations to explore the pattern of self-efficacy and knowledge results. Our *t*-tests showed no significant group differences in self-efficacy as a function of simulator or expertise ($p > .05$). We also determined that high self-efficacy was associated with quicker post-test completion time for SynDaver trainees ($r = -.58$, $p < .05$), but not for Laerdal trainees ($r = -.04$, $p > .05$).

Recently, Gignac and Zajenkowski (2020) advised additional analyses to test the existence of miscalibrated confidence. In their article, they recommend two methods to confirm the presence of a Dunning-Kruger effect, defined as the tendency for those unskilled at a task to be unaware of their own incompetence (Kruger & Dunning, 1999). Following Gignac and Zajenkowski's recommendations, we discovered evidence supporting the presence of the Dunning-Kruger effect in our training sample. First, we conducted the Glejser (1969) test of heteroskedasticity (i.e., calculating residuals from the simple linear model 'self-efficacy ~ posttest completion time' and correlating them with the posttest procedure completion time; $r = .35$, $p < .05$. These findings provide evidence of heteroskedasticity. Higher completion times were associated with greater residuals in the model, suggesting those with higher ability levels were more consistent (between participants) in their self-efficacy ratings than those with lower ability (see Figure 10). Thus, Hypothesis 6 was supported.

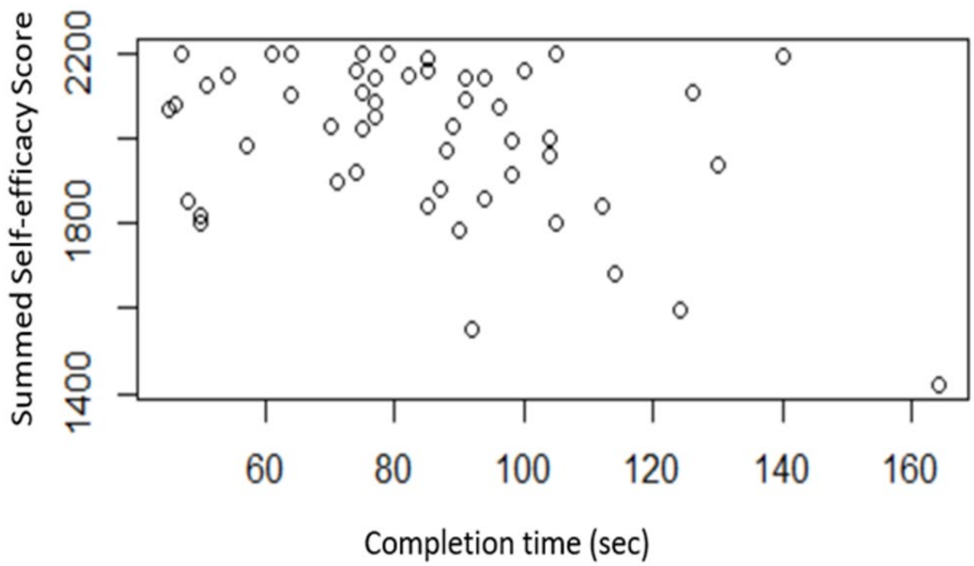


Figure 10. Scatterplot of the Relationship Between Objective Ability and Subjective Ability.

The second analysis suggested by Gignac and Zajenkowski (2020) was to model a quadratic regression model between subjective and objective estimations of ability. They proposed this test because they believed the strength of the relationship between subjective and objective ability level should increase with increasing levels of objective performance (i.e., a positive quadratic term). Therefore, we created a quadratic regression model that included a linear term (i.e., self-efficacy ~ posttest completion time + posttest completion time^2), finding the quadratic term to be non-significant (b = -.05, p = .053). However, when the maximum value of a quadratic relationship seems to occur at x = 0, it is acceptable to omit the linear term from a quadratic regression model. We did so because we expected procedure completion time would assume highest self-efficacy values when procedure completion time was at its smallest. After removing the linear term, the quadratic term was statistically significant (b = -.02, $p < .01$). This result provides evidence of a quadratic relationship, but in the opposite direction than proposed by Gignac and Zajenkowski. That is, our results suggested the relationship between subjective and objective ability level *decreased* with increasing objective ability level such that participants' subjective ability seemed to reach the ceiling effect of the measure and plateau at moderate objective ability levels (see Figure 11).

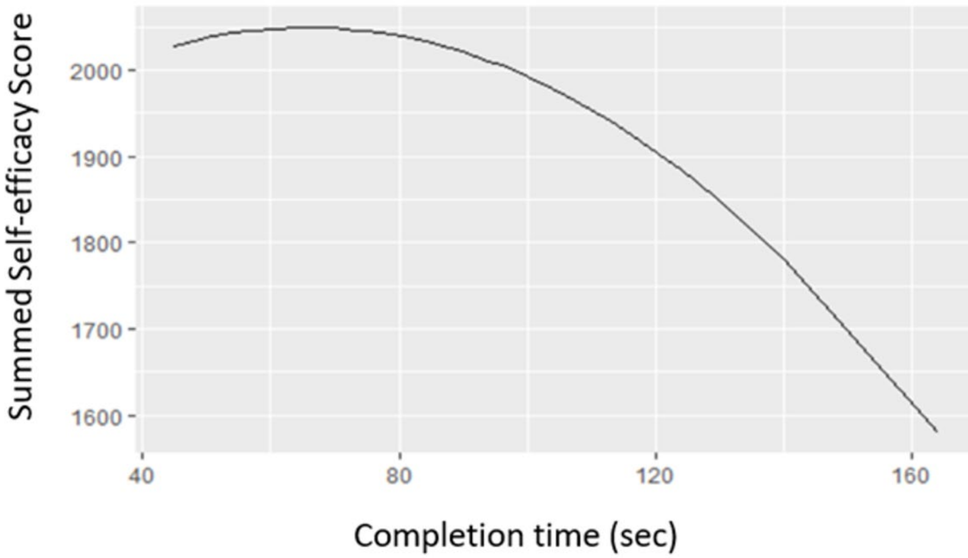


Figure 11. Quadratic Relationship Between Objective Ability and Subjective Ability.

Further exploratory analyses demonstrated task knowledge accounted for incremental variance in self-efficacy after controlling for performance ($\Delta R^2 = .08$, $p < .05$), and that skill acquisition factors (i.e., task experience, effort, and the simulator one trains on) have unique main and/or interactive effects when predicting Dunning-Kruger indices. The results of the Dunning-Kruger analyses on the Phase 2 data were written as a manuscript submitted for publication (Etherton, Bliss, & Winner, 2022).

We repeated the above analyses using omission and commission error rates as predictors. However, none of the analyses produced evidence of the Dunning-Kruger effect, perhaps due to the small levels of observed variability in the error scores.

Overall, our analyses demonstrated that simulator evaluation should consider confidence and competence calibration in training. Importantly, our results showed differences between trainees' calibration depending on which simulator they trained on.

3.4.4 Simulator Cue Ratings Hypotheses

We proposed three hypotheses related to simulator cue ratings. Hypothesis 7 stated that tactile cues would be rated as most realistic for the SynDaver trainer than for the Laerdal manikin. To test this hypothesis, we focused our analyses on the experienced participant group, because they have more applied experience with surgical cricothyroidotomy on actual patients. Considering those trainees, we examined the quantitative ratings data and the open-ended comments from the Simulator Cue Rating Worksheet. We tested this hypothesis using the non-parametric equivalent of the one-way analysis of variance (the Friedman's test), because the questionnaire ratings constituted ordinal data. For the SynDaver trainer, our Friedman's analysis showed that the experienced participants rated tactile cues as most realistic and visual cues as least realistic, $x^2(7) = 6.00$, $p = .05$. In contrast, the experienced participants who trained using the Laerdal trainer indicated no differences in realism across cue dimensions, $p > .05$.

Hypothesis 8 stated that visual cues would be rated most realistic for the Laerdal trainer. As noted above, experienced participants' ratings of visual cue realism were non-significantly different from their ratings of tactile and auditory cues, indicating that this hypothesis was not supported.

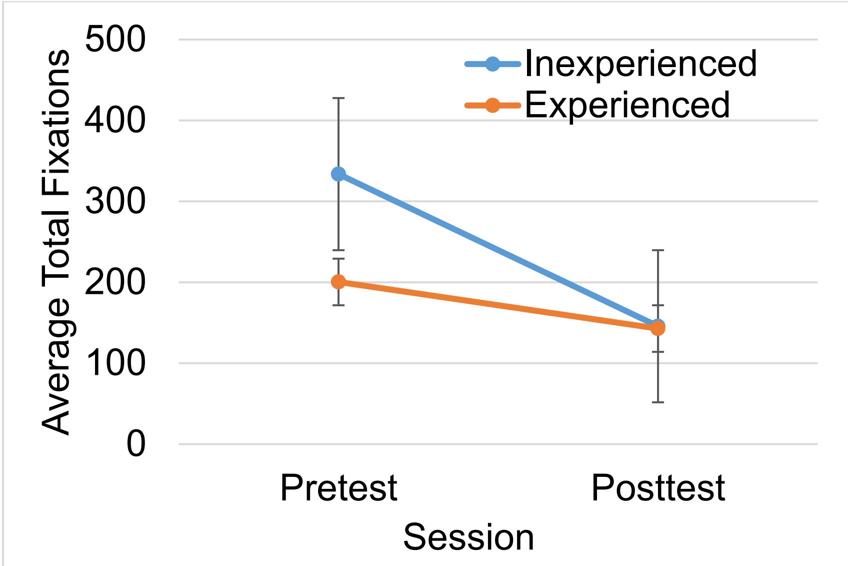
Hypothesis 9 stated that auditory cues would be rated as less realistic than tactile and visual cues. Likewise, this hypothesis was not supported because of the non-significant difference between cue responses.

To supplement the quantitative ratings, we analyzed the qualitative comments provided by experienced participants. To do so, we used a combination of latent semantic analysis and feature-related sentiment analysis (Liu, 2010) to determine the semantic themes represented within the comments. For the SynDaver trainer, latent semantic analysis indicated that the terms "membrane," "feel," and "structure" were mentioned most often (ten or more times). "Easier," "feel," and "simulate" were mentioned most often when discussing the Laerdal trainer (eight or more times). SynDaver comments generally focused on membrane tissues and tactile sensations offered by the simulator. Laerdal comments focused on ease of use, tactile sensations, and interactions between the simulator and the Cric-Key equipment used during the procedure.

Feature-Related Sentiment Analysis is generally used to extract targeted feature information from comments provided by experienced respondents. When used to interpret spontaneously generated prose, significant effort is needed to determine statement polarity (positive or negative comments), then to associate comments with specific features of an object or concept. Because our Simulator Cue Rating Worksheet was structured according to procedural steps indicated in the cognitive task analysis, the process of determining statement polarity and reference was straightforward. SynDaver positive comments related to simulated anatomical landmark realism; negative comments stressed unrealistic depth of simulated anatomical structures. Laerdal positive comments related to the ease of identifying landmarks; negative comments stressed unrealistic interactions with Cric-Key tools such as the Cric-Knife and the Melker/bougie introducer.

3.4.5 Eye Tracking Hypotheses

Hypothesis 10 stated that the frequency of recorded eye fixations would be greatest for the tracheal region, where the primary task was to be performed. To address this hypothesis, fixations were recorded by the Tobii Pro Lab software (version 1.111) using the "I-VT Attention" filter. That filter set the velocity threshold parameter to 100 degrees/second. Experimenters manually cycled through the recorded fixations to determine the target of each fixation. The average number of fixations was recorded only during the testing sessions (porcine pretest and posttest) to ensure that they reflected task-relevant attention that was not disrupted by didactic training with the SME, like the training trials were. The average number of total recorded fixations for the experienced participants was 200.42 and 142.83 during the posttest. In contrast, inexperienced participants averaged 333.71 fixations during the pretest and 139.14 during the posttest (see Figure 12). A mixed Session X Simulator X Expertise analysis of variance revealed a significant Session X Expertise interaction, $F(1, 36) = 6.33$, $p = .016$, partial $\eta^2 = .15$, observed power = .69. Main effects of Simulator and Expertise were not significant ($p > .05$); however, the main effect of Session was significant, $F(1, 36) = 20.84$, $p < .001$, partial $\eta^2 = .37$, observed power = .993.



Note: Error bars represent standard error

Figure 12. Average Total Fixations During Pre-test and Post-test.

Table 5 below shows the recorded proportions of fixations according to the eleven areas of interest specified prior to data analysis. As shown in the table, the proportion of fixations for the tracheal area was more than twice as great as the proportions for any other area. This supported Hypothesis 10.

Table 5. Proportions of Eye Fixations Occurring in Predefined Areas of Interest.

Area of Interest	Proportion of Experienced Pretest Fixations	Proportion of Inexperienced Pretest Fixations	Proportion of Experienced Posttest Fixations	Proportion of Inexperienced Posttest Fixations
1. Bag-Valve-Mask	0.05	0.05	0.03	0.05
2 Bag-Valve-Mask Tube	0.05	0.04	0.02	0.01
3. Cric-Key Knife	0.13	0.09	0.06	0.05
4. Melker/Bougie Introducer	0.15	0.15	0.09	0.12
5. Stability Strap	0.02	0.03	0.01	0.01
6. Cuff Inflation Syringe	0.05	0.04	0.05	0.06
7. Stability Wedge	0.01	0.01	0.00	0.00
8. Task Board	0.08	0.08	0.16	0.13
9. Porcine Trachea	**0.32**	**0.31**	**0.36**	**0.36**
10. Tool Bag	0.07	0.05	0.06	0.06
11. Other	0.09	0.13	0.16	0.15

Hypothesis 11 stated that the total recorded eye fixations would decrease from pre-test to posttest. To test this hypothesis, we conducted a paired samples t-test, finding evidence of a significant decrease in the number of total fixations from pretest to post-test ($t(39) = 5.89$, $p < .01$). Thus, Hypothesis 11 was supported.

Though we made no formal hypothesis about the area of first focus, during training sessions participants were instructed to focus first on the Cric-Key tools. To determine whether this indeed occurred, we examined the frequency of first fixations on the Cric-Key tools. Indeed, in both the novice and expert samples, participants' first area of fixation was most frequently the Cric-Key tools.

Our final analysis using fixation data involved computing correlations between total fixation counts for the pretest and posttest with conventional performance data (procedure times and tabulated errors). Table 6 below indicates the significant correlations that were observed.

Table 6. Significant Correlations Between Eye Fixations and Performance Variables.

Variable 1	Variable 2	Correlation
Pretest Total Fixations	Procedure Completion Time Differences	$r = .814$, $p < .001$
Pretest Total Fixations	Pretest Completion Times	$r = .821$, $p < .001$
Pretest Total Fixations	Pretest Omission Errors	$r = .546$, $p < .001$
Pretest Total Fixations	Pretest Commission Errors	$r = .410$, $p = .009$
Posttest Total Fixations	Pretest Procedure Completion Times	$r = .353$, $p = .025$
Posttest Total Fixations	Posttest Procedure Completion Times	$r = .729$, $p < .001$
Posttest Total Fixations	Pretest Commission Errors	$r = .425$, $p = .006$

3.4.6 Eye Tracking Scan Path Efficiency Analysis

No hypotheses were developed for eye tracking scan sequencing because of the expected difficulty of anticipating eye scan patterns for novices as they learned the task. The experimenters approached eye tracking scan sequences by manually reviewing each of the Tobii videos to determine the sequence of interest areas exhibited by participants during the pretest, posttest and five training sessions. Because some Tobii video recordings were unusable due to participants looking under the glasses or improper calibration, the experimenters first reduced the inexperienced and experienced data sets to ensure that the videos were interpretable. This reduction resulted in $N = 11$ participants in each experience group (22 total). Therefore, the experimenters reviewed 154 video files.

After reducing the video files, the experimenters consulted the cognitive task analysis and the SME to determine the optimal sequence of eye movements that should have occurred during the surgical cricothyroidotomy task. The optimal sequence specified the following visual targets, in order:

1. Tools/Tool Bag
2. Introducer/Syringe
3. Trachea
4. Cric-Knife
5. Trachea/hook
6. Introducer
7. Syringe
8. BVM/Tube

The first two steps (visually scanning the tools, introducer, and syringe) were critical for testing the equipment prior to beginning the procedure. This step is indicated as a critical aspect within the TCCC procedure list; however, it is rare to see it acknowledged within published research related to evaluation of simulator effectiveness. The third step (scanning the trachea) was important for identifying anatomical landmarks. The fourth step (scanning the Cric-Knife) was necessary when performing the incisions. The fifth step (scanning the trachea again or the hook) occurred when the participant used either a finger or the hook to widen the incision in preparation for introducer entry (either could be used). The sixth step (scanning the introducer) was necessary to insert it within the tracheal incision hole. The seventh step (scanning the syringe) occurred when the participant inflated the endotracheal balloon to stabilize the introducer within the trachea. The final step (scanning the bag-valve-mask and/or tube) accompanied manual ventilation. In addition to support from the cognitive task analysis, the sequence indicated was also taught to participants during each of the five training sessions.

To establish interrater reliability, two experimenters independently reviewed a subset of eight inexperienced and experienced participants' video sets (56 recordings), coding the eye scan sequences they observed. Because of the complexity of the procedure and potential for alternative procedural elements (e.g., using one's finger instead of the hook to widen the incision hole, using the bag-valve-mask (BVM) without the tube for manual ventilation, etc.), the experimenters adopted coding rules to ensure consistency of interpretation. Those rules are as follows:

- The set of tools was initially considered to be a single focal point when they were initially removed from the tool bag.
- The syringe and introducer were considered a single focal point ("syringe") during initial equipment testing.
- The Cric-Knife and trachea were considered to be singular focal points during incision, though participants frequently alternated between them repeatedly.
- The Cric-Knife and hook were considered separate focal points only after the hook was removed from the holder within the Cric-Knife.

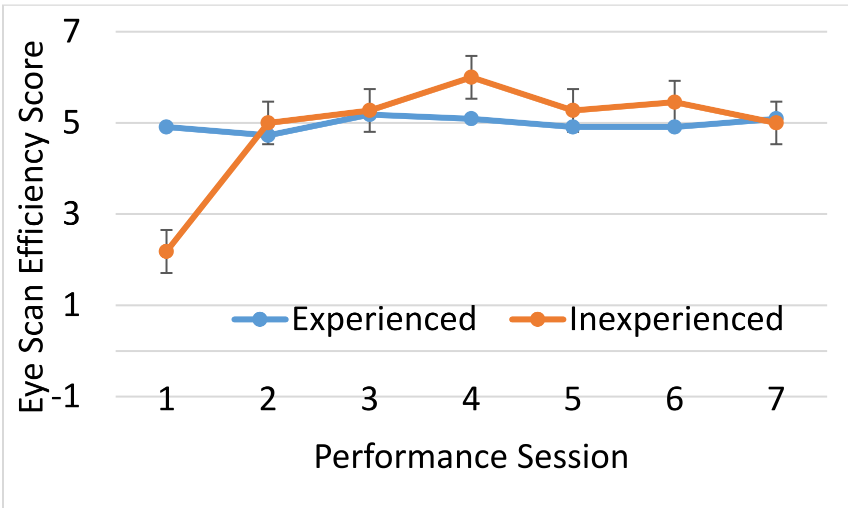
- Focal points were disregarded when the task involved physical, tactile manipulation. For example, inserting one's finger into the incision hole or pulling the bougie from the introducer.
- The bougie was not considered a separate focal point because it was part of the introducer until removed.
- Ambiguity of focal points was generally resolved by considering the task context and ongoing physical actions). For example, participants often looked at overlapping objects during tool testing. To resolve this ambiguity, raters referred to the object that was being held.

Following manual review of the 56 video files, each of the documented sequences was assigned a numeric score following the algorithm specified in Table 7 below. The algorithm assigned point values for certain eye scan sequences. This practice was derived from the cognitive task analysis and approved by our subject matter expert. After scoring each of the 56 videos, scores between the two raters were correlated for interrater reliability. The resulting coefficient was $r(56) = .62$, indicating moderate interrater reliability. The score was then calculated for the rest of the 154 videos.

Table 7. Elements of the Scoring Algorithm Applied During Eye Scan Efficiency Coding.

Observed Behavior	**Rationale**	**Points Awarded**
Looking at the tools **before** looking at the trachea	Equipment Testing	1
Looking at the syringe **before** looking at the trachea	Equipment Testing	1
Looking at the trachea **before** looking at the Cric-Key knife	Anatomical Landmark Identification	1
Looking at the hook (or trachea) **before** looking at the introducer	Widening Incision Hole to Ensure Tracheal Entry	1
Looking at the syringe (again) before using the bag-valve-mask	Stabilizing Introducer within Trachea	1
Efficiency Bonus:		
- Procedure Completion in 8 to 10 scan steps		2
- Procedure Completion in 11 to 13 scan steps		1
- Procedure Completion in more than 13 scan steps		0
Total Possible Point Score:		7

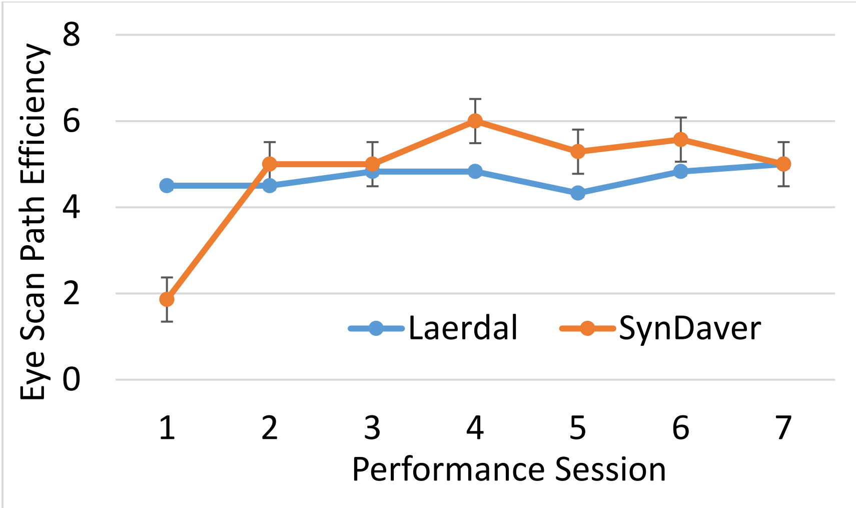
After calculating the derived eye scan sequence efficiency score for each of the 154 videos in the master set (77 inexperienced, 77 experienced), the results were analyzed using a 7 (session) X 2 (experience) X 2 (simulator) mixed-factor analysis of variance. The results indicated no omnibus interaction; however, there was an interaction of experience and session for efficiency, $F(6,108) = 8.58$, $p < .001$, partial $\eta^2 = .32$, observed power = 1.00; this was likely due to the significant difference in score during the pre-test session, with inexperienced participants showing significantly less eye scan efficiency (mean score = 2.18) than experienced participants (mean score = 4.9) (see Figure 13). There was also a significant effect of experimental session, $F(6, 108) = 9.8$, $p < .001$, partial $\eta^2 = .35$, observed power = 1.00. That effect consisted of linear, $F(1,18) = 16.52$, $p = .001$, partial $\eta^2 = .48$, observed power = .97; quadratic, $F(1,18) = 37.38$, $p < .001$, partial $\eta^2 = .68$, observed power = 1.00; and cubic, $F(1,18) = 11.98$, $p = .003$, partial $\eta^2 = .40$, observed power = .91, components (see Figure 13).



Note: Error bars refer to standard error.

Figure 13. Eye Scan Efficiency Score as a Function of Participant Experience and Experimental Session.

The between-groups main effect for experience was not significant ($p > .05$); however, there was a significant main effect for simulator, $F(1,18) = 4.64$, $p < .05$, partial $\eta^2 = .21$, observed power = .53. Examination of the means revealed that participants who used the SynDaver simulator exhibited greater average eye scan path efficiency than those who used the Laerdal simulator (see Figure 14).



Note: Error bars refer to standard error.

Figure 14. Eye Scan Efficiency Score as a Function of Simulator Used and Experimental Session.

During the eye scan efficiency data coding process, experimenters noted the following observations:

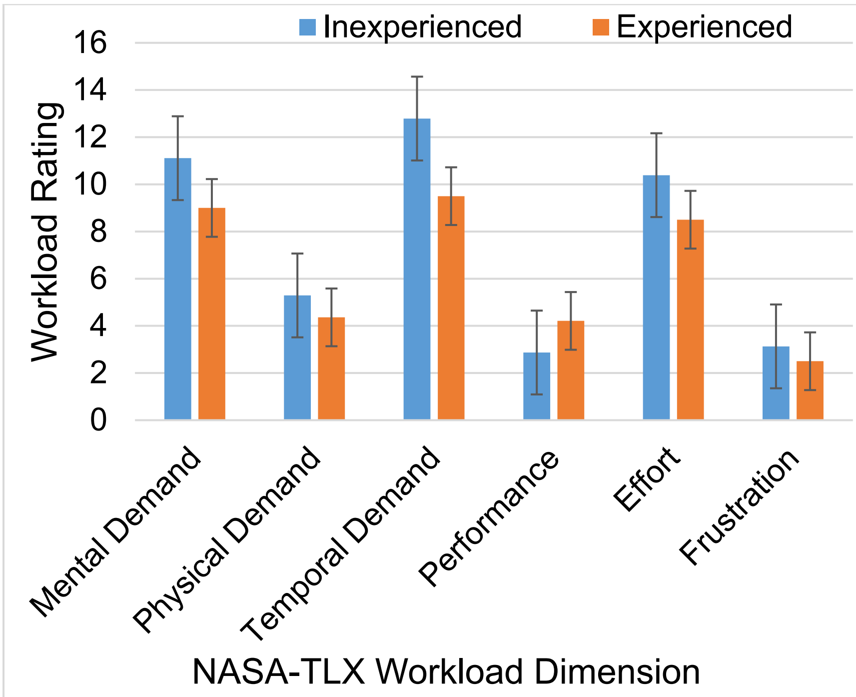
- Often, participants would anticipate their next physical move with their eyes (example: they would look at the BVM while using the syringe, before picking it up).
- Tactile task elements would sometimes cause participants to look forward/up so their fixation is not on the task (example: pulling the bougie from the introducer, they might look at the experimenter).
- In general, determining point of gaze during incision (knife or trachea) was very difficult; in many cases, the choice was an arbitrary one. For that reason, raters adopted an approach that relied on physical manipulation to resolve ambiguities.
- The configuration of the hook within the Cric-Knife was awkward for some participants.

The eye scan efficiency hypothesis testing demonstrated the value of eye-tracking metrics in surgical evaluation contexts. The results of this effort were presented at the 2022 Department of Defense Human Factors and Ergonomics Technical Advisory Group (Bliss, Etherton, Hodge, & Winner, 2022b).

4.4.7 Cognitive Workload

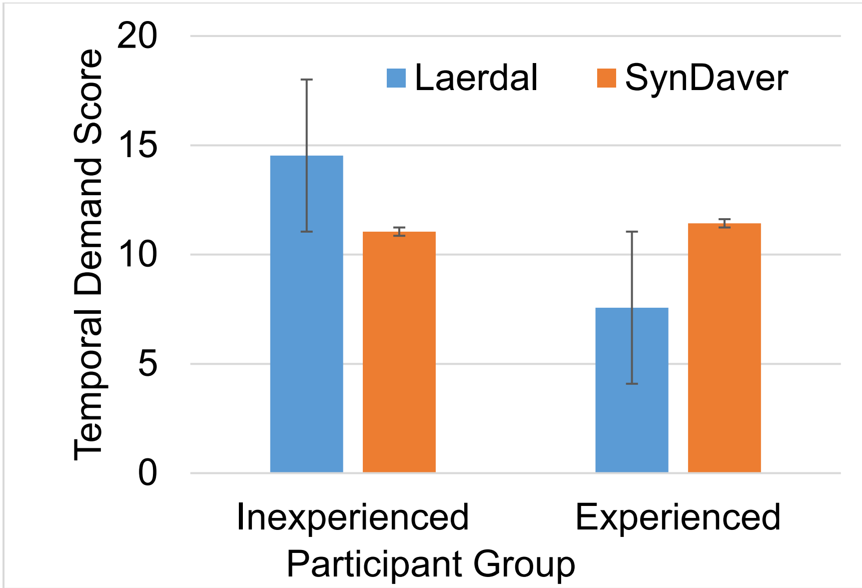
Following training, we administered the NASA-TLX questionnaire to participants to determine the level of cognitive workload they experienced while completing the training sessions. Because of the multidimensionality of the workload measure, we did not make any specific hypotheses regarding those data.

The observed data, separated by workload dimension, are illustrated in Figure 15. A three-way (simulator X experience X dimension) mixed analysis of variance revealed no significant interactions ($p > .05$). Similarly, main effects for simulator and expertise groups were not significant. However, the main effect for workload dimension showed that ratings were significantly different across dimensions, $F(4.07, 195.20) = 46.33$, $p < .001$, partial $\eta^2 = .49$, observed power = 1.00. A follow-up univariate analysis of variance for temporal demand showed a significant experience X simulator interaction (see Figure 16). Inexperienced trainees faced greater time demand than experienced trainees when using the Laerdal trainer, $F(1, 48) = 7.52$, $p = .009$, partial $\eta^2 = .14$, observed power = .77. This was manifested by a significant main effect of experience, $F(1, 48) = 6.06$, $p = .017$, partial $\eta^2 = .11$, observed power = .67.



Note: Error bars represent standard error.

Figure 15. NASA-TLX Summed Workload Rating by Expertise and Workload Dimension.



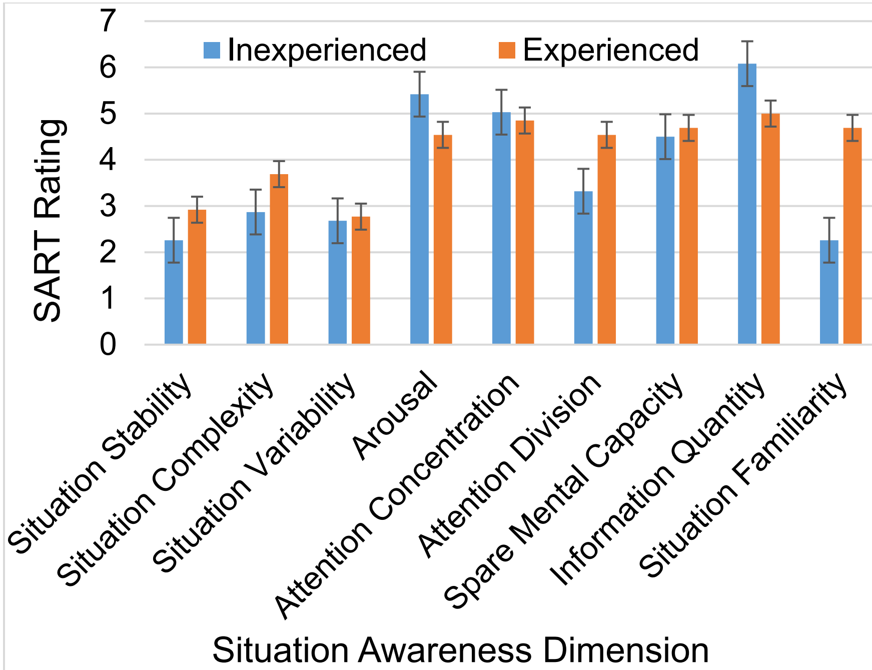
Note: Error bars represent standard error.

Figure 16. NASA-TLX Temporal Demand Rating by Expertise and Simulator Used.

3.4.8 Situation Awareness (SA)

Similar to our approach for cognitive workload, we also administered the SA Rating Technique (SART) questionnaire after training. The intent of the questionnaire was to assess participants' levels of situation awareness during training across several dimensions. Given the construct-related complexity of the measure, we did not make specific hypotheses about the data from the SART questionnaire.

Figure 17 illustrates the SA data as a function of situation awareness dimension, simulator used, and participant experience level. The 2 (experience) X 2 (simulator) X 9 (situation awareness dimension) interaction was significant, $F(5.17, 242.80) = 2.30$, $p < .05$. We computed a follow-up experience X situation awareness dimension interaction, $F(5.17, 242.80) = 2.30$, $p < .05$, partial $\eta^2 = .10$. We also noted a significant simulator X situation awareness dimension interaction, $F(5.17, 242.80) = 3.18$, $p < .01$, partial $\eta^2 = .06$, observed power = .89. Main effects analyses indicated that SART ratings significantly differed across dimensions, $F(5.17, 242.80) = 18.24$, $p < .001$, partial $\eta^2 = .28$, observed power = 1.00. We completed a targeted Kruskal-Wallis nonparametric test, showing that the SA construct of Situation Familiarity was rated significantly higher by experts than by novices, $\chi^2(1) = 14.31$, asymptotic significance < .001.



Note: Error bars represent standard error.

Figure 17. SA Rating by Experience and Situation Awareness Dimension.

3.4.9 Correlations among Measures

In addition to the analyses of variables discussed above for hypothesis testing, we also computed correlations among variables to determine observed relationships among constructs during the experiment. Those correlations are presented in Table 8.

Table 8. Significant Correlations among Dependent Variables

Variable 1	Variable 2	Correlation
Summed Self-Efficacy Score	Cognitive Test Score	-.30, $p < .05$
Summed Self-Efficacy Score	**Posttest Completion Time**	**-.58, $p < .01$**
Summed Self-Efficacy Score	Summed Simulator Cue Rating	.31, $p < .05$
Pretest Completion Time	Pretest Commission Errors	.32, $p < .05$
Pretest Completion Time	Posttest Commission Errors	.25, $p < .05$
Pretest Completion Time	Pretest Omission Errors	.33, $p < .05$
Posttest Completion Time	Posttest Commission Errors	.25, $p < .05$
Pretest Omission Errors	Pretest Commission Errors	.39, $p < .05$
Pretest Omission Errors	Pretest Completion Time	.42, $p < .05$
Pretest Eye Fixation Count	**Pre-post Completion Time Difference**	**.81, $p < .01$**
Pretest Eye Fixation Count	**Pretest Completion Time**	**.82, $p < .01$**

Pretest Eye Fixation Count	**Pretest Omission Errors**	**.55, $p < .01$**
Pretest Eye Fixation Count	**Pretest Commission Errors**	**.41, $p < .01$**
Posttest Eye Fixation Count	Pretest Commission Errors	.35, $p < .05$
Posttest Eye Fixation Count	**Posttest Completion Time**	**.73, $p < .01$**
Posttest Eye Fixation Count	**Posttest Commission Errors**	**.43, $p < .01$**
Summed Tactile Cue Ratings	Cognitive Test Score	-.28, $p < .05$
Summed Visual Cue Ratings	Pre-post Completion Time Difference	.32, $p < .05$
Summed Visual Cue Ratings	Posttest Completion Time	-.35, $p < .05$
Summed Auditory Cue Ratings	Pretest Omission Errors	-.28, $p < .05$
Posttest-Pretest Completion Time Difference	**Pretest Completion Time**	**.94, $p < .01$**
Posttest-Pretest Completion Time Difference	Pretest Omission Errors	.32, $p < .05$
Posttest-Pretest Completion Time Difference	Pretest Commission Errors	.29, $p < .05$
Post-Performance Score	Pretest Omission Errors	-.29, $p < .05$

Note: Boldfaced correlations significant at $p < .01$ level)

3.5 Phase 3 Method

Phase 3 consisted of three primary efforts: continued questionnaire data collection, a literature review on cricothyroidotomy simulation, and completion of a missing simulator cues analysis.

3.5.1 Phase 3 Method – Questionnaire Data Collection

In the final phase of our research, we initially intended to expand our human performance data collection approach to include experienced combat medics. However, our research team was forced to abandon physical data collection plans in early 2022 as COVID-19 infection rates spiked across the US. During that period, access to military facilities was restricted and ability to procure needed simulator components was jeopardized. Consequently, the team pivoted, choosing to extend our examination of simulator fidelity by distributing relevant questionnaires to applied military medical training experts with firsthand knowledge of simulators used in the field. Our rationale for this approach was grounded in the knowledge that adoption and use of CEMs for applied surgical cricothyroidotomy training to military personnel is influenced by the growth of technologies, cost of simulation platforms, and perceived ability of the CEMs to functionally represent target procedures. Because of increased popularity of advanced technologies such as augmented reality (AR), virtual reality (VR), and 3D printing, we considered it important to allow respondents to comment about the utility and fidelity of such approaches.

It is common for simulator designers to include the potential for scenario execution at many different levels. Only by experiencing such augmentations can trainees experience the full spectrum of patient reactions. Often, the simulator is used to supplement classroom training. In such situations, it can be challenging to assess the degree to which acquisition of skill is attributable to the simulator or to the broader pedagogical experience.

To complement our data collection activities during the first two phases, we constructed three questionnaires for distribution to military medical training schoolhouses across the Department of Defense. The first questionnaire included a background/demographics questionnaire that included items pertinent to the trainees, curriculum, CEM most often used, and instructors associated with each schoolhouse. The second questionnaire was a generalizable “fidelity perceptions” questionnaire that had items that were not simulator-specific that assessed trainees’ perceptions about the capability of the simulation device to approximate realistic task performance. The third questionnaire was a replication of the Simulator Cue Rating Worksheet used during the first two phases of data collection. This final questionnaire allowed ratings and comments to be provided that were relevant to the specific training simulator most often used at the schoolhouse. Our initial intent was to supplement our first two data collection phases by obtaining a substantial number of questionnaire responses from applied training practitioners from applied DoD training schoolhouses. By collecting detailed data from the questionnaires described above, our examination was intended to offer unique impressions of the fidelity of available simulator technologies currently being used to train surgical cricothyroidotomy and other combat trauma care procedures.

The analytic plan was centered around the employment of targeted questionnaires to reflect complex elements of the umbrella construct of simulator fidelity. Sub-constructs of greatest interest included physical and functional fidelity associated with certain simulator technologies such as extended reality and three-dimensional printing. Also of interest was the general relation between simulator functionality across task elements and the fidelity associated with each.

As a result of the COVID-inspired limitations, the research team devised an alternative data collection plan that would rely on questionnaire data to assess frequency, fidelity, and utility of cricothyroidotomy simulator use across DoD centers. Ideally, the questionnaire data would indicate how frequently centers made use of novel simulation technologies, notably three-dimensional and immersive simulation systems. Importantly, the anticipated respondents were to be administrative personnel who were responsible for procuring cricothyroidotomy training systems or for utilizing them as tools during training.

3.5.1.1 Questionnaire Data Collection Hypotheses

The research team did not propose formal hypotheses concerning the questionnaire data. Rather, given the plethora of developing technologies related to cricothyroidotomy simulation, the exercise was meant to be an exploration of impressions across the DoD regarding the adoption of simulation varieties. The hope was that results would shed light on the perceived advantages and disadvantages of established and proposed technologies. Those characteristics were to feature fidelity prominently, as it is the cornerstone of simulation efficacy.

3.5.1.2 Questionnaire Data Collection Measure Development

Acknowledging the importance of brevity for questionnaires, the research team focused development efforts on three questionnaires that would require no more than 45 minutes to complete. The three questionnaires were a background (demographics) questionnaire, a generalizable “fidelity perceptions” questionnaire that had items that were not simulator-specific, and the initial Simulator Cue Rating Worksheet from the in-person data collection effort completed in 2021. Instructions for the questionnaires included the directive to focus on the simulation platform that was used most often by the respondent's organization.

Draft versions of each of the three questionnaires indicated above were developed during spring 2022. The research team initially planned to implement them using a survey creation software package such as Qualtrics or SurveyMonkey. Unfortunately, none of the packages available met our criteria for cost and data security. Ultimately, we decided to construct .pdf versions of the surveys and e-mail them to participants using a government computer system. The AFRL Institutional Review Board approved this approach in June of 2022.

3.5.3 Phase 3 Method – Literature Review

The purpose of the literature review was to better understand the simulation landscape for the chosen example task of cricothyroidotomy. This task also addressed several of the identified validity inferences relevant to our fidelity evaluation method. The literature review focused on simulator comparisons for cricothyroidotomy, and general evaluations of fidelity associated with such simulators. The goals of the literature review were to summarize the work by outside teams in the scientific literature and to combine their conclusions with those that our team generated based on data collection and analysis activities.

The contractor team used PsychINFO and Google Scholar databases to find relevant articles with keywords such as 'cricothyroidotomy simulator' and 'Cric-Key method'.

3.5.4 Phase 3 Method – Missing Simulator Cue Analysis

The purpose of the missing simulator cue analysis was to document the participants' observations from the in-person data collection discussed in Phase 2, to associate them with likely cognitive implications, and to propose candidate solutions for simulator designers and military trainers to consider.

The missing cues were identified with a combination of qualitative observations from the participants that had trained with the simulator in Phase 2's data collection and using experimenter observations of the simulator. Each missing cue was associated with a sensory channel, a likely impact on cognition from its absence, and possible solution/remedy as proposed by the contractor team.

3.6 Phase 3 Results

Phase 3 was particularly important to rigorously approach the question of training device fidelity. Reflecting that, our contributions during this phase are as follows:

- Quantification of the impact of fidelity on military decision makers tasked with training technology procurement.
- Identification of technologies that represent extremes (low and high) of fidelity when used with trainees possessing little to no prior task experience.
- Indications of future tendencies of military decision makers to embrace novel CEM technologies and priorities of such individuals as they address anticipated training needs.

3.6.1 Phase 3 Results – Questionnaire Data Collection

Following the analyses conducted with in-person performance data, the research team's intent was to collect more fidelity perceptions data, focusing on newer technologies for cricothyroidotomy simulation, including simulators that utilize immersive displays and simulators that feature 3D printed components. Mixed, virtual, augmented, and extended reality technology have remained popular for medical simulation training, with several designers and manufacturers creating systems that replicate many of the capabilities of traditional animal models and simulators. Often, AR medical simulators allow virtual information to be overlaid onto a physical manikin for guidance and training purposes. Alternatively, VR trainers completely immerse the learner into the simulated experience. Immersive solutions have been used for guidance, team practice and communication and facilitate repetitive practice without the need for new materials. Lui et al. (2005) described an early computer model that allowed the participant to see a virtual patient and interact with a menu to palpate, incise, widen the incision, and intubate. The extent of bleeding varied depending on the depth of the cut. Comparable technology evaluated more recently by Sankaranarayanan et al. (2022) used force sensors under a palpation interface that allowed users to hold a haptic device and choose tools to complete cricothyroidotomy. The system then automatically recorded performance metrics for assessment. Another model with VR capability was the CricSim VR system, which used a virtual patient for cricothyroidotomy training. The system showed a 3D image of a patient's head, torso, and neck. The patient presented with a gunshot wound and CICO conditions. CricSim operated with stereoscopic 3D imaging through use of stereographic shutter glasses and haptic VR provided by two haptic interface devices (Proctor & Campbell-Wynn, 2014).

In addition to immersive simulations, some trainers have utilized simulators with 3D printed components. 3D tracheas are increasingly used for cricothyroidotomy. Multiple recent studies have utilized a free, online available 3D printing model for a trachea (Calvo et al., 2021; Huang, 2021; Kei et al., 2019;). The model was initially provided by Duggan et al. (2017); the online file is free of charge at the airway collaboration website (http://www.airwaycollaboration.org/3d-cric-trainer-1). Other researchers have created and utilized their own models; Katayama et al. (2019) created a novel 3D-printed trachea based off tomography scans. Other materials used in conjunction with the 3D models have included animal meat or skin, layered silicone pads (e.g., Dragon Skin, EcoFlex), self-adhering medical wrap (Coban), and Durapore surgical tape. Many solutions incorporate silicone because it resembles the texture and elasticity of actual human skin. In some cases, 3D solutions may feature the capability to bleed or breathe. An example is a simulator evaluated by Shaw and Hughes (2020). Perhaps the greatest benefits associated with 3D-printed models are low cost and improved accessibility. Simulators may range in cost from around $20 to $70 per unit.

The research team first performed a review of military centers, noting which were using VR/AR simulators for cricothyroidotomy training. We constructed an initial list of 32 expert contacts across the DoD and contracting agencies. Following our contact request, 14 of the contacts responded. We learned valuable information about a variety of systems; however, no experts indicated that the DoD had officially fielded any immersive systems for cricothyroidotomy training. A secondary review was then conducted of the specific systems that existed. That review identified 13 systems as indicated in Table 9 below.

Table 9. AR/VR Based Systems for Cricothyroidotomy Simulation.

System Name	System Manufacturer	DoD Adoption Status
AUGMED	DesignInteractive, Inc.	Not Adopted
PERSIM	MedCognition, Inc.	Adopted at March Air Reserve Base and William Beaumont Army Medical Center
System for Telementoring with Augmented Reality (STAR)	Purdue University	Not Adopted
Airway Lab	Arch Virtual, Inc.	Not Adopted
Virtual Airway Skills Trainer (VAST)	Rennsalaer Polytechnic Institute	Not Adopted
IMMERSE	ONR, CHI Systems, Inc. and George Mason University	Not Adopted
Integrated Virtual Augmentation System (IVAS)	Christoph Leuze, Andreas Zoellner, Alexander Schmidt, Robin Cushing, Marc Fischer, Kristin Joltes, & Gary Zientara (2021)	Not Adopted
C3ARESYS (Combat Casualty Care Augmented Reality Intelligent Training System)	SoarTech, Inc.	Not Adopted
Virtual Patient Immersive Trainer (VPIT)	Unveil Systems, Inc.	Not Adopted
Advanced Responsive Tactically Effective Military Imaging Spectrometer (ARTEMIS)	Univ. California – San Diego with Visual Arts, M. Maryland and Naval Medical Center San Diego.	Not Adopted
SimX	SimX, Inc.	Adopted by the Veteran's Administration
Cricothyroidotomy Emergency Procedure Training Simulator (CEPTS)	Val G. Hemming Simulation Center, Virtual Medical Environments Laboratory	Not Adopted
Engineering & Computer Simulations (ECS)	Mayo Clinic (Jacksonville)	Not Adopted

3.6.1.1 Respondent Recruitment

Following approval of the proposed questionnaire research approach by the AFRL IRB, the research team began to recruit potential respondents. The target demographic included military or civilian individuals who were responsible for procuring and/or using simulators to train military medical force members to perform surgical cricothyroidotomy according to the steps recommended by the TCCC committee. We began the recruitment process by constructing a list of contact emails from our SME, Retired Colonel Douglas Hodge. He initially provided a list of 99 emails that represented organizations and individuals responsible for training surgical cricothyroidotomy to military members. The organizations contacted included the following:

- US Army
    - 68W Combat Medical Specialist
    - 18D Special Forces Medical Sergeant
- US Navy
    - 8494 Marine Corps Combat Medical Corpsman
    - Marine Corps East
    - Marine Corps West
- US Air Force
    - Independent Duty Medical Technician (IDMT)
- US Joint Forces
    - Joint PA Program, Joint Base San Antonio
    - Defense Medical Readiness Training Institute (DMRTI) Joint Medical Readiness Training
- US Department of Health Administration

In addition to that initial list, the team accessed a list of email addresses from the National Association of Emergency Medical Technicians (NAEMT) web site. That list included 148 contact emails that represented military training personnel associated with military institutions. Last, our subject matter expert constructed an additional list of 56 contacts. In all, we sent emails to 293 individuals. As respondents returned the questionnaires, the team coded the data in Microsoft Excel according to predefined conventions established before data collection. Ultimately, we received data from $N = 17$ respondents, representing a response rate of 6%.

3.6.1.2 Results

Respondents ranged from 0 to 29 years of military experience, with an average of 15.8 years. Five respondents represented the U.S. Air Force, nine represented the U.S. Army, and five represented other categories, including National Guard units. Two indicated multiple organizations. All respondents indicated current TCCC certification, 13 indicated Advanced Cardio Life Support certification, and 12 indicated Pediatric Advanced Life Support certification. Nine indicated status as a medical doctor.

All respondents indicated that they were responsible for training individuals. Almost half (8) indicated that they were responsible for purchasing training technology. Nine indicated that they evaluated simulators as part of their role. Six indicated that they maintained simulators. All but one of the respondents indicated teaching open (surgical) cricothyroidotomy with the Cric-Key tool set; this matches the technique we taught participants in our initial experiment during 2021. All but one taught cricothyroidotomy monthly or quarterly; the remaining respondent indicated teaching it yearly. Sixteen of 18 respondents indicated that their students were new task learners; the remaining two indicated either refresher or mastery training. Only two respondents indicated that their students had prior applied practice with cricothyroidotomy; the rest were a mix of didactic familiarization or no experience at all. All but one indicated a requirement for learners to complete didactic training prior to simulator practice.

Eight respondents indicated using manikin simulators to train surgical cricothyroidotomy (two Simulaids, two Trauma FX, one Laerdal, one Caesar Tactical Operation Medical Manikin (TOMM), one Strategic Operations 6-in-1 trainer, and one unspecified "manikin"). Six respondents indicated they most frequently taught surgical cricothyroidotomy using an animal model (five of the six used porcine tracheas; one used sheep tracheas). Two used 3-D printed models, and two used human (perfused cadaver) models. None indicated using immersive (AR or VR) simulators. Half of the respondents give qualitative feedback only to trainees; the others offer a mix of qualitative and quantitative feedback.

When asked about their simulator's most valuable feature, participants' responses included high realism (7), availability (1), expense (1), ability to use with real people (1), availability of anatomical landmarks (1), ease of parts replacement (3), ease of maintenance (1), and size/portability (1). Problematic features cited were low realism (9), inconvenience of procurement (4), cost (1), and inability to represent multiple training situations (1). Only one respondent indicated that the chosen simulator was "not as good" as a human model for reproducing the task (see Table 11).

Table 10. Simulator and Most Valuable and Problematic Features Reported.

Respondent	Simulator	Most Valuable Feature	Most Problematic Feature
1	Pig Trachea	Realism	Butchering schedule
2	Hog Tracheas	Availability and expense	Inner hog trachea membrane is not present on humans
3	Pig Trachea	Cric-Key kit	It is rubber
4	3-D Printed Model	We can place it on a real person	Made of hard plastic…too easy to find membrane
5	Swine trachea; lung and cadaver	Similarity to human anatomy	Still only swine anatomy
6	Manikin	Landmarks	Real feel
7	6 in 1 trainer (strategic operations)/ Pig/ TOMM trauma simulator	Easy to replace parts	Not realistic/gives very little feedback
8	Perfused cadavers	Realism	Procurement
9	Trauma FX Manikins	Replacement skins	Highly Identifiable Site

10	Laerdal	[none]	[none]
11	Simulaids Cric Trainer 101-135-GSA	Easy to maintain	Not as realistic as other options
12	Porcine tracheas	Realism of actual tissue	Obtaining it for each training
13	Trauma FX manikins	Easy to use and change out parts	It does not feel like real skin and depending on the model you can't get feedback from the BVM
14	Strategic Medical RT 3-D printed tracheas	Realistic	Cost
15	Simulaids Deluxe Cric Trainer	Small and portable	Not realistic, poor quality
16	Human model	Realism of human patient model	Limited number of procedures for each model
17	Animal (sheep) model	Animal models are more realistic	Acquiring, storing, transporting, and disposing of animal models

3.6.1.3 Fidelity Perceptions Questionnaire

Using the Fidelity Perceptions Scale, participants rated their most frequently used simulator on several items on a five-point graphic ratings scale. After coding the rating items, we adjusted for the reverse coding of items 5, 10, 11, 13, and 16. Original plans involved scale validation efforts for the scale, but due to our low response rate, we instead simply computed means for animal models, human models, manikin models, 3D printed models, and an overall mean. The means are listed in Table 12 below. Although all categories had few responses, the ratings fell in line with expectations. Notably, the average 3D printed model rating was higher than all others except the human models. Also, manikins had the lowest average fidelity rating.

Table 11. Mean Fidelity Perceptions Scale Fidelity Ratings for Identified Simulator Categories.

Simulator Category	Mean Fidelity Rating
Animal Model (*N*=6)	2.66
Human Model (*N*=2)	3.89
Manikin Model (*N*=7)	2.11
3D Printed Model (*N*=2)	3.47
Overall (*N*=17)	3.55

3.6.1.4 Simulator Cue Rating Worksheet (Quantitative Analyses)

Table 13 below shows the average rating value from 0 (Cue is not available from the simulator) to 4 (Simulator is extremely realistic for detection of this cue) for the different simulator categories. These data show that participants rated the animal models higher than other categories, with manikins rated lowest in fidelity.

Table 12. Mean Simulator Cue Rating Scale Ratings for Identified Simulator Categories.

Simulator Type	Mean Fidelity Rating
Animal Model (*N*=5)	3.13
Human Model (*N*=2)	3.01
Manikin Model (*N*=7)	2.94
3D Printed Model (*N*=2)	3.07
Overall (*N*=17)	2.97

3.6.1.5 Simulator Cue Rating Worksheet (Open Ended Responses).

In addition to the quantitative rating data, the Simulator Cue Rating Worksheet, we coded textual responses to each of the items. The responses, though based on a subset of the 17 participants, substantiated clear deficiencies of each simulator regarding physical cues that are necessary for training. Most notable were the lack of blood and mucus upon incision, which may lubricate the tracheal opening to facilitate bougie entry while obscuring certain anatomical features and tissues. Other common observations were inability of each trainer to support observation of cues related to transfer of air during ventilation (tube misting, chest sounds, bilateral chest rise). It was notable that inexperienced and experienced participants' opinions converged on these points. More detailed coverage of the observations is recorded within a separate report entitled "Missing Simulator Cue Cognitive Impact Analysis" (See Appendix C).

3.6.2 Phase 3 Results – Literature Review

The literature review on cricothyroidotomy simulation yielded many important findings (See Appendix B). Such findings were both with respect to how such literature reviews might be completed by simulation personnel in the future (for non-cricothyroidotomy use cases), but also for cricothyroidotomy-specific simulation findings. The literature review incorporated considerable information related to the investigation of readiness within experimental investigations and supported (or not) by available training devices for surgical cricothyroidotomy. The literature review also identified various needs for curriculum design for training airway procedures.

The general recommendations for conducting literature reviews to highlight fidelity evaluation had to do with the sources chosen, the time window, the weighing of information, aspects of review (i.e., selection and organization of relevant psychological constructs), peer review versus public access information.

Other, more specific recommendations and findings included the following:

- The highest fidelity simulator may not necessarily be the most effective simulator for training. Many lower-fidelity simulator options have been favorably rated for skill acquisition. Depending on training goals, using a lower fidelity simulator may allow learners to quickly acquire task fundamentals with minimal expenditure of time or funds.
- Though a high-fidelity simulator often facilitates performance improvements, lower-fidelity simulators have been shown to achieve the same goal. In fact, in nearly all the simulator training studies, all models used for training let to some improvement of skill.
- Though study authors or simulator developers may consider their models to be “high-fidelity,” such characterizations do not necessarily equate to faithful representation of the target task. The construct of fidelity has been defined in a plethora of ways, leading to confusion in the scientific and lay training communities.
- When considering simulators, it is important to hold important psychological fidelity. Specifically, it is critical to determine whether the simulator prompts appropriate responses throughout the simulation. This relates to the presentation of proper cues *in situ*.
- Ensuring that there are opportunities for feedback is critical for simulation-based training. Failure to do so will jeopardize training effectiveness and learner comprehension.
- Opportunities for repetitive practice are necessary for muscle memory and precise skill development, especially for novice learners.

3.6.3 Phase 3 Results – Missing Simulator Cue Analysis

The Missing Simulator Cue Analysis document identified several key takeaways with respect to simulator evaluation (See Appendix C). It detailed the associated sensory channels, cognitive impacts, and proposed resolutions for multiple missing cues in both Laerdal SIMMAN ALS and SynDaver Adult Cricothyroidotomy Trainer simulators.

One finding from the analysis was that most missing cues for the two cricothyroidotomy simulators were visual or tactile in nature. This is partially attributable to the nature of the task of cricothyroidotomy and highlights how the types of cues that may be present vs missing in a simulator vary in modality between task environments. That is, some tasks may be more auditorily demanding (e.g., communications-heavy positions like joint terminal attack controller; JTAC) and require a closer examination at the auditory cues present in simulation environments.

Another recurring theme of the results of the analysis was that Augmented Reality and Virtual Reality solutions were proposed as possible solutions to compensate for the detrimental effect of missing certain cues. This highlights the importance of understanding such emerging technologies such that they might augment the fidelity of simulators.

Overall, the Missing Simulator Cue Analysis directly compared the physical cues present in the distal simulators (i.e., Laerdal and SynDaver) to those present in the real-world task (i.e., live patient cricothyroidotomy) and yielded unique insights to the physical fidelity of the two simulators.

4.0 OVERALL RESULTS

Following a multiphasic plan, we explored variations in physical and functional fidelity associated with common, off-the-shelf human patient simulation systems. We based our determination on available, published estimates of device fidelity for the procedure. Following completion of a cognitive task analysis informed by medical experts (emergency medical technicians with recent applied experience performing surgical cricothyroidotomy) we recruited research participants to learn and perform cricothyroidotomy using one of the simulator systems under investigation. As they did so, we recorded task completion success, speed, accuracy, efficiency data, measures of eye fixation and scan path efficiency, and levels of situation awareness and cognitive workload. After collecting performance data from experienced participants, we replicated our data collection process by training inexperienced trainees (medical students who lacked training in surgical cricothyroidotomy). By documenting and comparing their performances, we documented the potential for the selected training simulators to support skill acquisition. Our final activities included administration of a series of fidelity questionnaires to applied training personnel, completion of a cricothyroidotomy simulation literature review, and completion of a missing cues analysis on the two cricothyroidotomy simulators we used in the in-person data collection phase (Phase 2). The resulting data documented subjective and objective metrics of simulation fidelity currently in use at military medical institutions.

After data was collected, the contractor team compiled the notable findings from each of the data sources that supported various validity inferences. As a reminder, supporting validity inferences is distinct from supporting hypotheses; a supported validity inference implies the inference was informed by some form of data collection (either qualitative or quantitative), without any claim whether the inference was determined to be true or not. Overall, the information gathered in support of the validity inferences provided valuable insights that collectively inform a fidelity evaluation.

Inference 1 stated the distal simulator performance domain is representative of performing cricothyroidotomy on a distal (i.e., non-porcine) simulator. This inference was supported through the cognitive task analysis by identifying the necessary behaviors, in sequence, with time expectations per behavior to successfully perform cricothyroidotomy using the Cric-Key method. Because this method of prescribed behaviors, sequences, and timing is applicable to simulators and live patients alike, the method was deemed appropriate when gauging success of cricothyroidotomy using a distal (i.e., non-porcine) simulator. Thus, Inference 1 was supported for both Laerdal and SynDaver simulators.

Inference 2 stated the proximal simulator performance domain is representative of performing cricothyroidotomy on a proximal (i.e., porcine) simulator. Similarly, this inference was supported with the cognitive task analysis of cricothyroidotomy using the Cric-Key method. Because this method of prescribed behaviors, sequences, and timing is applicable to simulators and live patients alike, the method was deemed appropriate when gauging success of cricothyroidotomy using a proximal (i.e., porcine) simulator. Thus, Inference 2 was supported.

Inference 3 stated the real-world task performance domain is representative of performing cricothyroidotomy on a live patient. Again, this inference was supported with the cognitive task analysis of cricothyroidotomy using the Cric-Key method. Because this method of prescribed behaviors, sequences, and timing is applicable to simulators and live patients alike, the method was deemed appropriate when gauging success of cricothyroidotomy with a live patient. Further, the task analysis identified additional behaviors that are specific to live patient contexts. For example, there were six preliminary behaviors identified as critical before deciding to begin a cricothyroidotomy. Such behaviors included determining an ability to ventilate or intubate the patient, consideration of non-medical causes (e.g., abnormal anatomy, anterior airway, superior thyroid cartilage), and consideration of additional accommodations due to patient body type (e.g., consider hyperextending neck if patient is obese). Also unique to the real-world task performance domain, the number of physical cues present on a live patient are obviously more numerous and realistic than on any simulator. Further, the outcomes of cricothyroidotomy using the Cric-Key method are different when completing the task on a live patient; the objective of the task is the same (i.e., induce an artificial airway) but with an additional superordinate goal of preserving human life that isn't present in simulated environments. With the completion of the cognitive task analysis, Inference 3 was supported.

Inference 4 stated cues, behaviors, and outcomes associated with successful task performance are all interrelated. This inference addresses the assumption that cues, behaviors, and outcomes are interrelated and informed by one another, as proposed by our fidelity taxonomy. This assumption was supported by the linking of objectives (i.e., induce an artificial airway) with behaviors (e.g., use the cric knife with dominant hand to incise a vertical incision…) and cues (e.g., feel for structure using non-dominant hand) as part of the cognitive task analysis. Thus, Inference 4 was supported.

Inference 5 stated the extent to which cricothyroidotomy using the distal (i.e., non-porcine) and proximal (i.e., porcine) simulators are related. This inference was supported by a significant relationship between performance measures from the proximal (i.e., porcine) simulator and both distal simulators: Laerdal ($r = .60$, $p < .01$) and SynDaver ($r = .43$, $p < .05$). The significant relationship indicates performance on the distal simulator is predictive of performance on the proximal simulator. Thus, Inference 5 was supported.

Inference 6 reflected the extent to which cricothyroidotomy using the proximal (i.e., porcine) simulator and live patient are related. In our literature review, we discovered studies that provided empirical support for the claim that porcine trachea simulators were perceived as more realistic than synthetic rubber trachea models (Cho et al., 2008). Also, porcine tracheas may not be as standardized in shape or structure as synthetic models (Mabry et al., 2015), but such variability is representative of the variability in live patient cricothyroidotomy. As such, there was justification for porcine simulators being considered a close representation of live patient cricothyroidotomy, supporting Inference 6.

Inference 7 stated the extent to which cricothyroidotomy using the distal (i.e., non-porcine) simulator and live patient is similar. We collected multiple data sources to support this inference. Some of the data sources provided direct support whereas others provided indirect support. One data source that provided direct support of Inference 7 was the simulator cue ratings which assessed the physical fidelity between distal simulator and real-world task through quantitative ratings of cue existence and realism. Further, a qualitative data source providing direct support was the missing cue analysis that identified which physical cues were missing in either distal simulator (i.e., Laerdal and SynDaver) relative to the real-world task (i.e., live patient cricothyroidotomy), the cues' sensory channel, cognitive impact, and potential resolutions/remedies. A data source providing indirect empirical support of Inference 7 was the collection of training performance data to first examine the relationship between performance using the distal and proximal simulators and then support through qualitative information (i.e., literature review, cognitive task analysis) the assumption that the proximal simulator is representative of the real-world task. Such indirect methods of supporting validity inferences are necessary when the real-world task is not measurable; that is, it is impossible to truly replicate live patient cricothyroidotomy so the value of any significant relationship in performance between two simulators is contingent on support that one of the simulators is a close approximation of the real-world task. Thus, there were multiple data collection tools used to satisfy the assumptions of fidelity evaluation (i.e., Inference 7).

Inference 8 stated the chosen measure of performance (i.e., completion time) for the distal (i.e., non-porcine) simulator adequately represents the behaviors necessary for performance. The cognitive task analysis identified the criteria for successful task completion: the necessary behaviors, sequence, and timing of such behaviors. As such, a critical criterion for performance in this time-sensitive task was completion time. The team used completion time as the operationalized variables representing the behavioral aspects of the proximal and distal simulator performance domains. Thus, Inference 8 was supported.

Inference 9 stated the observed performance measures (i.e., completion time) between the distal and proximal simulators are significantly related. Indeed, the team tested the relationship between completion time on the distal simulator and the proximal simulator immediately after. The relationship between observed performance measures (i.e., completion time) for the distal and proximal simulators were statistically significant for both distal simulators: Laerdal ($r = .60$, $p < .01$) and SynDaver ($r = .43$, $p < .05$). Thus, Inference 9 was supported.

Inference 10 stated the chosen measure of performance for the proximal (i.e., porcine) simulator (i.e., completion time) adequately represents the behaviors necessary for performance. Like stated for Inference 5, Inference 7 was supported by the cognitive task analysis through identification of completion time as a critical metric of cricothyroidotomy success. Thus, Inference 10 was supported.

Inference 11 stated the chosen measure of performance for the proximal (i.e., porcine) simulator (i.e., completion time) is representative of live patient cricothyroidotomy's performance domain. The cognitive task analysis determined that an appropriate metric of successful live patient cricothyroidotomy is in accurately performing the procedure as quickly as possible to minimize the risk of losing the patient's life or causing permanent brain damage. As such, using completion time as an operational definition of the proximal simulator's performance domain was reasonable, supporting Inference 11.

Inferences 12 and 13 pertained to the simulator cue rating worksheet. After having trained with a simulator, respondents were given the simulator cue rating worksheet, which listed all necessary cues for live patient cricothyroidotomy (as determined by a cognitive task analysis). Respondents were asked to rate whether each cue existed and its degree of realism in the simulator. Inference 12 stated the simulator cue ratings measure assesses the perceived availability and salience of cues in the distal (i.e., non-porcine) simulator. This inference was supported when respondents were instructed to fill out the simulator cue rating worksheet in reference to the non-porcine simulator they had trained with. Inference 13 stated the cues listed on the simulator cue ratings measure are the necessary cues for live patient cricothyroidotomy. This was supported by defining the list of cues present in the measure by using the list of cues present in the real-world task (i.e., live patient cricothyroidotomy). Thus, Inferences 12 and 13 were supported.

The validity inferences listed above were all supported with information gleaned from various data sources; as such, the relevant assumptions in fidelity evaluation research have been addressed and a comprehensive view of the simulators' fidelity may be assessed with greater confidence.

Using all the data sources mentioned previously, the team was able to quantify several elements of both simulators being examined (i.e., Laerdal and SynDaver). We ran analyses within and between the two distal simulators; that is, evaluating the fidelity as if either simulator was being examined in isolation (i.e., within) and running comparative analyses between the two (i.e., between). Table 14 displays an abbreviated summation of the quantitative metrics and analyses that resulted from the fidelity evaluation method.

Table 13. Fidelity Evaluation Quantitative Summary.

Data Source	Simulator Fidelity Findings (Within)		Simulator Fidelity Findings (Between)
	Laerdal SIMMAN ALS	SynDaver Adult Cricothyroidotomy Trainer	Laerdal & SynDaver
Experimental Performance Data (Phase 2)	**Performance on Laerdal was significantly related to performance on the porcine model** (i.e., Relationship of completion time of Laerdal on final practice trial and porcine trial immediately after ($r = .60$, $p < .01$))	**Performance on SynDaver was significantly related to performance on the porcine model** (i.e., Relationship of completion time of SynDaver on final practice trial and porcine trial immediately after ($r = .43$, $p < .05$))	**The relationship between distal simulator performance and proximal simulator performance was similar between Laerdal and SynDaver** (i.e., interaction term of simulator X distal simulator completion time predicting proximal simulator completion time; $b = -.10$, $p = .82$)
	Training on Laerdal reduced procedure completion time (Change in porcine completion time from pre- to post-test ($t(25) = 9.75$, $p < .01$))	**Training on SynDaver reduced procedure completion time** (Change in porcine completion time from pre- to post-test ($t(25) = 7.46$, $p < .01$))	**Training on either Laerdal or SynDaver produced similar improvements in procedure completion time over time** (i.e., effect of simulator in a repeated measures ANOVA, $F(1) = 2.37$, $p = .13$)
	Training on Laerdal reduced the frequency of commission errors (Change in porcine commission errors pre- to post-test ($t(25) = 5.70$, $p < .01$))	**Training on SynDaver reduced the frequency of commission errors** (Change in porcine commission errors pre- to post-test ($t(25) = 6.96$, $p < .01$))	**Training on either Laerdal or SynDaver produced similar decreases in commission errors over time** (i.e., effect of simulator in a mixed factorial ANOVA, $F(1) = 0.00$, $p = 1.00$)
	Training on Laerdal reduced the frequency of omission errors (Change in porcine omission errors pre- to post-test ($t(25) = 6.37$, $p < .01$))	**Training on SynDaver reduced the frequency of omission errors** (Change in porcine omission errors pre- to post-test ($t(25) = 6.65$, $p < .05$)	**Training on either Laerdal or SynDaver produced similar decreases in omission errors over time** (i.e., effect of simulator in a mixed factorial ANOVA, $F(1) = 1.38$, $p = .25$)
Simulator Cue Rating Worksheet (Phase 2)	**On average, respondents rated Laerdal cues between "Simulator is considerably realistic for detection of this cue" and "Simulator is extremely realistic for detection of this cue"** ($M = 3.24$, $SD = .51$)	**On average, respondents rated SynDaver cues between "Simulator is considerably realistic for detection of this cue" and "Simulator is extremely realistic for detection of this cue"** ($M = 3.11$, $SD = .38$)	**There was no significant difference in cue ratings between Laerdal and SynDaver trainees** (All cues: $U = 285.00$, $p = .33$; Tactile cues: $U = 304.00$, $p = .53$; Visual cues: $U = 259.00$, $p = .15$; Auditory cues: $U = 311.50$, $p = .62$)
Missing Cue Analysis (Phase 3)	**There were 14 missing cues identified in the Laerdal simulator**	**There were 13 missing cues identified in the SynDaver simulator**	**n/a**

5.0 OVERALL DISCUSSION AND CONCLUSIONS

5.1 Fidelity Evaluation Conclusions

As previously mentioned, it is important to consider fidelity evaluation holistically, as an overall view of multiple perspectives. That is, from Table 14 there is no single analysis that can be considered the singular test of fidelity. Rather, there are several sources of data and analyses that, when used together, provide a more comprehensive evaluation. When examining the fidelity of cricothyroidotomy simulators, results suggested three overall findings: (1) training on either Laerdal or SynDaver simulators led to similar improvements in performance, (2) perceived fidelity for both Laerdal and SynDaver simulators was comparably high, and (3) both Laerdal and SynDaver simulators had many missing cues. Because the various validity inferences were supported, the team can more confidently state both Laerdal and SynDaver simulators are effective at training cricothyroidotomy and there appears no evidence that one is significantly superior to the other along any of the performance metrics tested. However, despite non-significant differences between the collected performance metrics, there were other notable quantitative outcomes of performance on either simulator that were highlighted in Phase 2 Results (Section 3.4). Also, the current method generated considerable qualitative data that should be considered before making purchasing decisions. For example, as identified in the missing cue analysis (Appendix C), the Laerdal simulator possessed a physical forehead, which indicates proper patient orientation. If trainers are particularly interested in training trainees how to identify and correct proper head orientation for the procedure, such training would be impossible using the SynDaver simulator. Even if there are no significant differences between the two simulators with performance or perceptual data, some of the qualitative differences may be critical difference-makers when deciding whether to purchase the simulator.

Despite non-significant differences in performance metrics between the two examined distal simulators, this method demonstrated a more comprehensive view of fidelity along a greater number of metrics than prior methods. It is common for simulator developers to claim their simulator is “high fidelity”, without providing empirical support for such a claim. Unfortunately, this contributes to the preexisting issue of definition confusion when it comes to the term ‘fidelity’. What fidelity is and how it can be measured is a topic of great importance, and even greater confusion. With the method developed and tested in this task order, we demonstrated how the collection and analysis of empirical data can provide a richer defense or critique of a simulator’s fidelity. The theoretical framework (presented in Figures 1 and 2) and the simulator evaluation method was prepared as a theoretically focused manuscript currently in preparation (Etherton, Bliss, & Winner, 2023).

The results of the current task order demonstrated a highly rigorous examination of simulation fidelity; however, many personnel do not have the time or resources to conduct the extensive data collection done herein. As such, we developed a method for assessing simulator fidelity that relies on minimized data availability (Appendix D). Such a method should be of value for simulator procurement officers who are tasked with selecting training devices. Current procedures often reflect the importance of trainer cost and functionality, rather than considerations of realism (c.f., Rooney, 2018; Ruisi, Wier, & Ziegenfuss, 2017). Consideration of such information sources should increase the efficiency of the simulator evaluation process.

5.2 Overall Conclusions

Combat casualty care is a particularly challenging mission for medics and other medically trained personnel. In addition to the technical challenges associated with case management, environmental, and battle related stressors necessitate unwavering mission focus, situation awareness and attention to detail. Such qualities clearly differ at an individual level and must be determined individually. To successfully ensure proficiency-centric readiness evaluation, simulation technologies must be capable of faithfully representing task elements. Ensuring that deficiencies are noted and quantified is essential to support effective personalized learning and may enable future refinement of human patient simulators to react intelligently in response to learner performance gaps.

The research described in this report is important to the USAF and DoD for several reasons. It extends prior work within RHWL (RHA) that focused on providing systematic, competency-based fidelity methods for training environments. Accomplishing this goal improves the robustness and adaptability of military training operations. Though rarely practiced, integrated learning monitoring and assessment is critical to future deployment training prioritization. Fidelity research informs decision making by medical forces concerning training technology acquisition. It also enables medical personnel performance to be linked to larger force mission readiness. Among other organizations, the Defense Medical Modeling & Simulation Office (DMMS) and the Air Education and Training Command (AETC) have each expressed great interest in the conduct of simulation fidelity evaluation and optimization to ensure the convenience, power, and success of medical trauma combat casualty care.

Historically, personnel training has relied on the conventional “see one/do one/teach one” philosophy of task mastery discussed by Miller (1990). Human patient simulators have supplemented training procedures and have recently been emphasized by the military as a supplement (or at times, a replacement) for classroom training (Dorlac, Bishop, & Dorlac, 2014). As the USAF replaces the Self Aid and Buddy Care program with the TCCC program, ambitious goals (such as the training of 3,000 Airmen every 18 months) will rely strongly on CEM’s (Air Force Surgeon General Public Affairs, 2019) and on rapid learning that is personalized for each trainee. The research conducted here scrutinizes the fidelity and capability of representative CEM’s to help optimize personalization of medic education and training.

One of the primary reasons for collecting data across experimental phases was to investigate potential simulator comparison approaches based on fidelity assessment. Most results revealed that the simulators used were comparable in terms of their ability to train the surgical cricothyroidotomy procedure. In fact, consideration of traditional performance measures (procedure completion times, subjective comments) suggested few differences between the trainers.

Assessment of fidelity proved to be a complex undertaking. Development of the Simulator Cue Rating Worksheet for the initial human performance experiment was helpful for allowing participants to express their opinions about simulator realism in a detailed, organized fashion. The numeric data confirmed the general inability of both simulators to represent certain cues associated with the presence of a full body (chest sounds, bilateral chest rise, etc.). More useful, however, were the open-ended response data, which indicated more subtle differences in tissue texture, anatomical feature representation, and landmark salience between the two simulators.

For this reason, we believe open-ended questions should be included on all future assessments of physical and functional fidelity to facilitate nuanced appraisals of realism.

We believe our results demonstrate the utility of cognitive task analyses, training efficacy methods and evaluations using both behavioral performance and self-reported fidelity perceptions data (as suggested by Bell & Waag, 1998), missing cue identification, and comprehensive literature review as part of a larger simulation fidelity evaluation method.

Assessments of skill acquisition during the first experiment revealed few surprises. As expected, participants gained expertise in the surgical cricothyroidotomy procedure, with their speed and accuracy performances appearing to stabilize within five training sessions (the number recommended in the literature). More noteworthy, however, were the noticeable performance differences that occurred between the final training session and the transfer post-test. Longer procedure completion times and fewer errors of omission and commission suggest that stabilized simulator performance may not predict efficacy on an animal model. Such a discrepancy indicates both non-porcine simulators did not lead to perfect transfer of skills to the porcine simulator. Curiously, the training-test transfer pattern was not reflected in eye scan path efficiency, suggesting independence of task attention patterns and execution behaviors during the posttest. Though few observations were represented in the questionnaire data gathering exercise, participants seemed to indicate that immersive and 3D simulator technologies may represent effective alternatives for skill mastery training. Clearly more research is recommended in this area.

Our emphasis on assessments of learner subjective confidence as well as objective competence reflects a trend in the medical community to consider learner confidence as a surrogate measure of capability. The data we gathered, however, suggested the presence of miscalibrated confidence, with a heteroskedasticity in the relationship between subjectively assessed confidence and objective measures of task performance. Notably, our analytical method reflected the assessment approach recommended by Gignac and Zajenkowski (2020). By doing so, we identified unexpected trends illustrating that the strength of the relationship between subjective and objective ability level was lower as objective ability increased; that is, subjective ability rose sharply from low- to mid- ranges of objective ability and plateaued with little change between mid- to high-ranges of objective ability. Further, the variability of subjective ability levels decreased in those with faster performance. Both findings would have been overlooked if we relied only on the bivariate relationship between confidence and competence. Moving forward, we recommend assessments of self-efficacy after didactic training and after practical instruction to further clarify the relationships suggested here. Our findings and recommendations regarding the calibration between trainee confidence and competence were written up as a manuscript and are currently under review (Etherton, Bliss, & Winner, 2022).

As suggested above, traditional measures of task performance were instrumental to shed light on participants' development of surgical skill, transfer of learning, attention focus, and task efficiency. Yet, the comprehensibility of our construct measurement efforts is important to emphasize. Had our team focused only on traditional performance speed and accuracy indices, sophisticated conclusions about attention, efficiency, and error would not have been possible. In the spirit of triangulation, our goal was to thoroughly assess constructs by employing detailed taxonomy-driven error analyses, latent semantic and feature-related sentiment analyses, eye fixation and scan path efficiency analyses, and follow-up questionnaire analyses to enable a comprehensive examination of performance and fidelity. The results of our efforts suggests that future investigators adopt a similar holistic measurement philosophy to achieve fidelity and performance assessment and discrimination.

In addition to the conventional assessment of constructs such as performance speed and accuracy, the research team also employed validated measures of workload and situation awareness. Analyses of the results from the first experiment yielded few novel insights; yet the asymmetric time demand influencing inexperienced trainees who trained with the Laerdal trainer may converge with comments noting the lack of tissue fidelity of that trainer. Situation awareness findings, meanwhile, revealed an expected difference in task familiarity between inexperienced and experienced participants. Given the diverse nature of the SART questionnaire and the observed interactions among conditions and groups, it is advisable to consider such constructs within future research efforts. It may also be worthwhile to consider similar constructs such as resilience and cognitive readiness, given the evident lack of learning transfer during the posttest.

The results reported here represent a foundational approach that may be replicated within future efforts. Recent advances in immersive technologies and 3D printing have expanded the possibilities for simulation design. Our questionnaire results suggest that not only have military medical institutions begun to adopt such technologies, but they have also taken proactive steps to fashion and implement hybrid solutions for task training. For example, emerging simulators have advocated creating 3D printed skin material for use on actual human models or in conjunction with physical or immersive simulation platforms. Such creative possibilities may reduce the cost of simulators, but they will surely make assessment of physical and functional fidelity more complicated.

Regardless of the forward path for fidelity research investigations, we expect the following conclusions to retain their importance:

- Cognitive Task Analysis – As noted throughout this report, cognitive task analysis represents the foundation of fidelity assessment. It is the organizational framework within which expert judgments are associated with sensory task requirements. It is also the most appropriate mechanism for determining alignment of training simulator features with required cues.
- Triangulation of Measures – Conventional measures of performance speed and accuracy are necessary but not sufficient for determining the influence of device fidelity. As clearly demonstrated here, the complexity of fidelity demands a broad assessment plan to distinguish attention from performance; opinion from demonstration; and confidence from competence. Future designers must demand a variety of evidence from device manufacturers and marketers to justify characterizations of "high fidelity."
- Distinction of Competence from Confidence – Long assumed to be linear, our data suggest the relationship between learner performance and self-efficacy is not straightforward. Such complexity demands a holistic consideration of constructs, refinement of measurement tools, and representation of relationships.
- Transfer of Training – As demonstrated by our data, stability of task performance does not guarantee comparable transfer performance. For this reason, any suitable examination of simulator fidelity must incorporate

transfer of training assessment. Only by doing so will the criterion validity of a simulation device be understood.

- Consideration of Didactic Training Elements – Our results confirmed the receptiveness of learners to didactic instruction. Participants demonstrated reduced errors, improved procedure completion times, and improved eye fixation and scan path tendencies after a single training session. For this reason, any assessment of training device fidelity must be managed in concert with an evaluation of the didactic training procedures in place.

In the short term, we believe that our investigation will provide the scientific and operational communities valuable information to position and employ CEMs within the medic training curricula. In the long term, data collected as part of this task order should allow better informed strategies for personalizing combat medic training. Doing so will ultimately streamline skill acquisition and lower costs.

6.0 REFERENCES

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention and transfer of human motor skills. *Psychological Bulletin, 101(1)*, 41-74.

Air Force Surgeon General Public Affairs (2019). New training prepares Airmen to save lives. [online]. Retrieved July 26, 2019 from https://www.airforcemedicine.af.mil/News/Display/Article/1820423/new-training-prepares-airmen-to-save-lives/

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents*, *5*(307-337).

Baum, D. R., Smith, D. A., Hirshfeld, S. F., Klein, G. A., Swezey, R. W., & Hays, R. T. (1982). *Specification of training simulator fidelity: A research plan*. (Technical Report No. 558). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *International Journal of Aviation Psychology, 8*, 223-242.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478.

Bliss, J.P., Etherton, K., Hodge, D., & Winner, J. (2022a). Performance based evaluation of cricothyroidotomy simulators for training. *Proceedings of the 2021 Human Factors and Ergonomics Society Health Care Symposium.* New Orleans, LA: March 22.

Bliss, J.P., Etherton, K.C., Hodge, D., & Winner, J. (2022b). *Using eye tracking metrics to quantify simulator learning by experienced and inexperienced medical trainees.* Poster presented at the 2022 Department of Defense Human Factors and Ergonomics Technical Advisory Group. Oklahoma City, OK: May 16.

Bliss, J.P., Etherton, K.C., Hodge, D., Winner, J., Millwater, T., McCool, J., & Cloude, E. (2021). *Applying error analysis to support surgical cricothyroidotomy simulator evaluation.* Poster presented at the 2021 Human Factors and Ergonomics Society Annual Meeting. Baltimore, MD: November 2021.

Bredmose P. P., Habig K., Davies G., Grier G., Lockey D. J. (2010). Scenario based outdoor simulation in pre-hospital trauma care using a simple mannequin model. Scan J Trauma Resusc Emerg Med, 18(13), 1-6.

Calvo, A., Esteve, C. I., Varela, V., Gomez-Lopez, L., Perdomo, J. M., Berge, R. & Sancho, C. G. (2021). Design, Application and Evaluation of a Cricothyrotomy Model for a Multidisciplinary Simulation. An Observational Single Centre Study. *EducaciónMédica, 22*(4)*,* 305-310. http://dx.doi.org/10.1016/j.edumed.2020.12.003

Carey, J. M., & Rossler, K. (2022). The how when why of high fidelity simulation. In *StatPearls [Internet]*. StatPearls Publishing.

Center for Army Lessons Learned (2017). Tactical combat casualty care: Lessons and best practices. Leavenworth, KS: Center for Army Lessons Learned.

Cho, J., Kang, G. H., Oh, Y. M., Choi, H. J., Im, T. H., Yang, J. H., Cho, Y. S., & Chung, H. S. (2008). Comparison of Manikin Versus Porcine Models in Cricothyrotomy Procedure

Training. *Emergency Medicine Journal, 25*(11), 732-734. https://dx.doi.org/10.1136/emj.2008.059014

Cronbach, L. J. (1970). *Essentials of psychological testing*. (3rd edition). New York: Harper & Row.

Crouch, L. L. (2010). High-fidelity human patient simulation experiences and baccalaureate nursing students' perceptions. *Clinical Simulation in Nursing, 6*(3), e111.

Demirel, D., Butler, K.L., Halic, T., Sankaranarayanan, G., Spindler, D., Cao, C., Patrusa, E., Molina, M., Jones, D.B., De, S., & deMoya, M. (2016). A hierarchical task analysis of cricothyroidotomy procedure for a virtual airway skills trainer (VAST) simulator. *American Journal of Surgery, 212*(3), 475-484.

Dorlac, W. C., Bishop, J. M., & Dorlac, G. R. (2014). Use of Simulation and military medical training: 2014 [online]. Retrieved July 26, 2019 from https://military-medicine.com/article/3124-use-of-simulation-military-medical-training-2014.html

Duggan, L. V., Lockhart, S. L., Romano, K. R., Weingart, S. D., Levitan, R. M. & Brindley, P. G. (2017). Front-of-neck airway meets front-of-neck simulation: improving cricothyroidotomy skills using a novel open-access three-dimensional model and the Airway App. *Can J Anesth, 64,* 1079-1081. http://dx.doi.org/10.1007/s12630-017-0926-9

Etherton, K., Bliss, J.P., Winner, J. (2022). *Examining Self-efficacy and its Miscalibration in Surgical Training Contexts*. Manuscript submitted for publication.

Etherton, K., Bliss, J.P., Winner, J. (2023). *Simulation Fidelity as a Form of Validation: Development of a Conceptual Model and Evaluation Method*. Manuscript in preparation.

Fautua, D. T., & Schatz, S. (2012). Cognitive readiness and the challenge of institutionalizing the "new" versus "news." *Journal of Cognitive Engineering and Decision Making, 6*(3), 276-298.

Feingold, C. E., Calaluce, M., & Kallen, M. A. (2004). Computerized patient model and simulated clinical experiences: Evaluation with baccalaureate nursing students. *Journal of Nursing Education, 43*(4), 156-163.

Gerathewohl, S. J. (1969). Fidelity of simulation and transfer of training: A review of the problem. Report AM 69-24. Department of transportation, Federal Aviation Administration, Office of Aviation Medicine, USA.

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80*, 101449. https://doi.org/10.1016/j.intell.2020.101449

Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, *64*(325), 316-323.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, *11*(3), 385.

Guion, R. M. (1987). Changing views for personnel selection research. *Personnel Psychology*, *40*(2), 199-213.

Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research"* (PDF). In P.A. Hancock & N. Meshkati (eds.), *Human Mental Workload*. Advances in Psychology. 52. Amsterdam: North Holland. pp. 139–183.

Hays, R. T., Singer, M. J. (1989). *Simulation fidelity in training system design: Bridging the gap between reality and training*. Springer-Verlag.

Hermens, F., Flin, R., & Ahmed, I. (2013). Eye movements in surgery: A literature review. *Journal of Eye Movement Research, 6*(4), 4.

Holding, D. H. (1965). *Principles of training*. London: Pergammon Press.

Holt, L., Schreiber, B., Schroeder, M., Call, L., Swigert, S., Greschke, D., Bennett Jr., W. (2012). *Evaluation and Improvement of the Deployable Tactical Trainer*. Report prepared for Air Force Research Laboratory, Wright-Patterson AFB, OH, USA.

Huang, J., Ocariz, S., Sims, C. R., Licatino, L. K., & Warner, P. A. (2021). A Novel Approach to Emergency Airway Simulation Using a 3D-printed Cricothyrotomy Task Trainer. *The Journal of Education in Perioperative Medicine, 23*(3), E670. http://dx.doi.org/10.46374/volxxiii_issue3_sims

Johnson, G. (2017). The relationship of fidelity on simulation performance. Unpublished doctoral dissertation, University of North Dakota.

Kahol, K., Vankipuram, M., & Smith, M. L. (2009). Cognitive simulators for medical education and training. *Journal of Biomedical Informatics, 42*(4), 593-604.

Katayama, A., Nakazawa, H., Tokumine, J., Lefor, A. K., Watanabe, K., Asao, T., & Yorozu, T. (2019). A High-Fidelity Simulator for Needle Cricothyroidotomy Training is Not Associated with Increased Proficiency Compared with Conventional Simulators. *Medicine, 98*(8), e14665. http://dx.doi.org/10.1097/MD.0000000000014665

Kei, J., Mebust, D. P., & Duggan, L. V. (2019). The Real Cric Trainer: Instructions for Building an Inexpensive, Realistic Cricothyrotomy Simulator with Skin and Tissue, Bleeding, and Flash of Air. *The Journal of Emergency Medicine, 56*(4), 426-430. https://doi.org/10.1016/j.jemermed.2018.12.023

Klein, G., & Militello, L. (2001). Some guidelines for conducting a cognitive task analysis. In *Advances in Human Performance and Cognitive Engineering Research* (vol. 1, pp. 161-199). Elsevier.

Krishnan, D. G., Keloth, A. V., & Ubedulla, S. (2017). Pros and cons of simulation in medical education: A review. *International Journal of Medical and Health Research, 3*(6), 84-87.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*(11), 1183.

Lapkin, S., & Levett-Jones, T. (2011). A cost–utility analysis of medium vs. high-fidelity human patient simulation manikins in nursing education. *Journal of Clinical Nursing, 20*, 3543-3552.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. Damerau (eds.), *Handbook of Natural Language Processing* (2[nd] ed).

Liu, A., Bhasin, Y. & Bowyer, M. (2005). A haptic-enabled simulator for cricothyroidotomy. *Studies in Health Technology and Informatics*, *111*, 308–3133.

Mabry, R., Frankfurt, A., Kharod, C., & Butler, F. (2015). Emergency Cricothyroidotomy in Tactical Combat Casualty Care. *Journal of Special Operation Medicine, 15*(3), 11-19. http://dx.doi.org/10.55460/KYGI-F0VF

Mabry, R. L., Nichols, M. C., Shiner, D. C., Bolleter, S., & Frankfurt, A. (2014). A comparison of two open surgical cricothyroidotomy techniques by military medics using a cadaver model. *Annals of Emergency Medicine, 63*(1), 1-5.

Mahmood, K. (2016). Do people overestimate their information literacy skills? A systematic review of empirical evidence on the Dunning-Kruger effect. *Communications in Information Literacy*, *10*(2), 3.

Maran, N. J., & Glavin, R. J. (2003). Low- to high-fidelity simulation: A continuum of medical education. *Medical Education, 37*, 22-28.

Mayo, P. H., Hackney, J. E., Mueck, J. T., Ribaudo, V., & Schneider, R. F. (2004). Achieving house staff competence in emergency airway management: Results of a teaching program using a computerized patient simulator. *Critical Care Medicine, 32*(12), 2422-2427.

Meyer, G. F., Wong, L. T., Timson, E., Perfect, P., & White, M. D. (2012). Objective fidelity evaluation in multisensory virtual environments: auditory cue fidelity in flight simulation.

Militello, L. G., & Hutton, R. J. B. (1998). Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics, 41*(11), 1618-1641.

Miller, R. (1954). Psychological considerations in the design of training equipment. (Final Report No. WADC-TR-54-563, AD71202). Dayton, OH: Wright Patterson Air Force Base.

Miller G. E. (1990). The assessment of clinical skills/competence/performance. Academic Medine, 65, S63-67.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Mund, E. (2011). Synthetic human tissue offers maximum realism. *EMSWorld*.

Osgood, C. E. (1949). The similarity paradox in human learning: a resolution. *Psychological Review, 56*(3), 132–143.

Pandian, V., Leeper, W. R., Jones, C., Pugh, K., Yenokyan, G., Bowyer, M., & Haut, E. R. (2020). Comparison of surgical cricothyroidotomy training: a randomized controlled trial of a swine model versus an animated robotic manikin model. *Trauma surgery & acute care open*, *5*(1), e000431.

Proctor, L. (2005). Simulation learning: Advancing medical education and patient safety through simulation learning. *Patient Safety & Quality Healthcare*, March/April.

Proctor, M. D. & Campbell-Wynn, L. (2014). Effectiveness, Usability, and Acceptability of Haptic-Enabled Virtual Reality and Mannequin Modality Simulators for Surgical Cricothyroidotomy. *Military Medicine, 178*(3), 260-264. http://dx.doi.org/10.7205/MILMED-D-13-00365

Prost, J. H., Schreiber, B. T., & Bennett Jr, W. (2007). *Identification and evaluation of simulator system deficiencies*. *Proceedings of the 2007 Interservice/Industry Training, Simulation and Education Conference (IITSEC), November 2009, Orlando FL*

Prost, J. H., Schreiber, B. T., Bennett Jr, W., & Kleinlein, K. (2008). Providing effective and efficient training: a model for comparing simulator improvements. *In Proceedings of the 2008 Fall Simulation Interoperability Workshop, September 2008, Orlando FL.*

Rehmann, A. J. (1995). *A handbook of flight simulation fidelity requirements for human factors research*. (Technical Report No. DOT/FAA/CT-TN95-46). Washington, D.C.: Department of Transportation.

Rhodes, L. M., & Curran, C. (2005). Use of the human patient simulator to teach clinical judgment skills in a baccalaureate nursing program. *Computer, Informatics, Nursing, 23*(5), 256-262.

Rodgers, D. L. (2007). High-fidelity patient simulation: A descriptive white paper report.Charleston, WV: Healthcare Simulation Strategies.

Rooney, D. M., Hananel, D. M., Covington, B. J., Dionise, P. L., Nykamp, M. T., Pederson, M., Sahloul, J. M., Vasquez, R., Seagull, F. J., Pinsky, H. M., Sweier, D. G., & Cooke, J. M. (2018). An approach to value-based simulator selection: The creation and evaluation of the Simulator Value Index tool. *Surgery, 163*, 927-932.

Roza, Z.C. (2004): *Simulation Fidelity Theory and Practice: A Unified Approach to Defining, Specifying and Measuring the Realism of Simulations.* Doctoral thesis: Delft University of Technology, Faculty of Aerospace Engineering, Delft.

Ruisi, J., Wier, G., & Ziegenfuss, A. (2017). Standardizing and validating medical simulation equipment procurement across the Air Force Medical Service. *Journal of Defense Modeling and Simulation: Applications, Methodology, and Technology, 14*(4), 421–426.

Sankaranarayanan, G., Odlozil, C. A., Hasan, S. S., Shabbir, R., Qi, D., Turkseven, M., De, S., Funk, G. & Weddle, R. J. (2022). Training on a Virtual Reality Cricothyroidotomy Simulator Improves Skills and Transfers to a Simulated Procedure. *Trauma Surgery & Acute Care Open, 7*, e000826. http://dx.doi.org/10.1136/tsaco-2021-000826

Scalese, R. J., Obeso, V. T., & Issenberg, S. B. (2007). Simulation technology for skills training and competency assessment in medical education. *Journal of General Internal Medicine, 23*(Supplement 1), 46-49.

Schmidt, R. A., & Bjork, R. A. (1989). *New conceptualizations of practice: Common principles in three research paradigms suggest important new concepts for practice*. (Draft Research Product). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Schmidt, R. A., & Bjork, R. A. (1992). *New conceptualizations of practice: Common principles in three research paradigms suggest important new concepts for training*. *Psychological Science, 3*(4), 207-217.

Selcon, S. J., Taylor, R. M. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations* (AGARD-CP-478) (pp. 5/1–5/8).

Shaw, M. R. & Hughes, K. E. (2020). High Risk, Low Volume: Evaluation of a Reusable Cricothyrotomy Model in a Paramedic Difficult Airway Training Course. *Air Medical Journal, 39*(5), 380-382. http://dx.doi.org/10.1016/j.amj.2020.05.017

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). Washington, DC; Author.

Swain, A. D., & Guttmann, H. E. (1983). Handbook of human reliability analysis with emphasis on nuclear power plant applications. (Technical Report). Washington, D.C.: U.S. Nuclear Regulatory Commission.

Tun, J. K., Alinier, G., Tang, J., & Kneebone, R. L. (2015). Redefining simulation fidelity for healthcare education. *Simulation and Gaming, 46*(2), 159-174.

Vreuls, D., & Obermayer, R. W. (1985). Human-system performance measurement in training simulators. *Human Factors*, 27(3), 241-250.

Winner, J., & Millwater, T.L. (2019). Evaluating human-patient simulation fidelity and effectiveness for combat-medical training. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*. Sage India: New Delhi, India: SAGE Publications.

7.0 LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

3D	Three Dimensional
AETC	Air Education and Training Command
AFRL	Air Force Research Summary
AFSC	Air Force Specialty Codes
ANOVA	Analysis of Variance
AR	Augmented Reality
ARTEMIS	Advanced Responsive Tactically Effective Military Imaging Spectrometer
BVM	Bag-Valve-Mask
C3ARESYS	Combat Casualty Care Augmented Reality Intelligent Training System
CAF	Combat Air Force
CALL	Center for Army Lessons Learned
CEM	Computer-Enhanced Mannequins
CEPTS	Cricothyroidotomy Emergency Procedure Training Simulator
CMRP	Comprehensive Medical Readiness Program
COVID	Corona Virus Disease
DMMS	Defense Medical Modeling & Simulation Office
DMRTI	Defense Medical Readiness Training Institute
DoD	Department of Defense
DoDHFETAG	Department of Defense Human Factors and Ergonomics Technical Advisory Group
DTT	Deployable Tactical Trainer
ECS	Engineering & Computer Simulations
FDA	Food and Drug Administration
HPS	Human Patient Simulators
HPW	Human Performance Wing
ICTL	Individual Critical Task List
IDMT	Independent Duty Medical Technician
IRB	Institutional Review Board
IVAS	Integrated Virtual Augmentation System
JTAC	Joint Terminal Attack Controller

LVC	Live, Virtual, Constructive
MARCH	Massive Hemorrhage, Airway, Respirations, Circulation, Head Injury/Hypothermia
NAEMT	National Association of Emergency Medical Technician
NASA-TLX	National Aeronautics and Space Administration - Task Load Index
PAW	Pain, Antibiotics, Wounds, Splinting
RH	Research-Human
SA	Situation Awareness
SART	SA Rating Technique
SART	Situation Awareness Rating Technique
SME	Subject Matter Expert
STAR	System for Telementoring with AR
TCCC	Tactical Combat Casualty Care
TOM	Tactical Operations Medical
TOMM	Tactical Operation Medical Manikin
USAF	US Air Force
USAFSAM	USAF School of Aerospace Medicine
VAST	Virtual Airway Skills Trainer
VPIT	Virtual Patient Immersive Trainer
VR	Virtual Reality
WPAFB	Wright-Patterson Air Force Base
WRT	Warfighter Readiness and Training

APPENDIX A - Cognitive Task Analysis

The following task analysis includes steps required to perform surgical cricothyroidotomy using the Cri-Key method. The steps that populate table cells were initially taken from task analysis articles published by Demirel et al. (2016) and Cannon-Bowers et al. (2013). However, because the Cri-Key method was not formally established and adopted until after those articles were published, the following represents an updated analysis specific to that procedure variant. Table entries were evaluated and refined by nine subject matter experts in surgical cricothyroidotomy.

Following the validation meeting with the subject matter experts, their comments and changes were considered and changes were implemented within the table where warranted. The resulting table, reproduced here, was used to design training materials, a sensory cue rating questionnaire, and protocol elements for the subsequent data collection sessions. The error elements of the table will be elemental for subsequent performance data analyses.

Task Analysis Table for Cri-Key enabled Surgical Cricothyroidotomy

Note: Task elements are indicated in the left column below. Items included in light gray are preparatory to the cricothyroidotomy procedure and will be introduced and discussed during the experimental training. However, they are not intended to be a part of the ultimate experimental task.

Task Element	Notes
A. Awareness of action: Get to patient	(Simulated patient available and accessible)
B. Assess: Evaluate: Inability to establish ventilate, place a sufficient rescue airway or inability to intubate	(Inability to intubate or ventilate assumed within scenario description)
C. Consider medical causes - Facial Trauma, Airway Burns, Block by Vomit or Aspirate, Foreign Body, Bronchical secretions	(No complicating conditions as stated in the scenario description)
D. Consider non-medical causes – Abnormal anatomy, anterior airway, superior thyroid cartilage, inferior or posterior cricoid cartilage, lateral cricoid and thyroid cartilage	(Patient anatomy is assumed typical within scenario description)
E. Locate and prepare equipment for	(Equipment provided and available next to simulated patient)

performing cricothyroidotomy	
F. Consider hyperextending neck if patient is obese	(Patient stature classified as normal in scenario description)

Task Element	**Time Required**	**Sensory Cues**	**Knowledge Required**	**Possible Errors**
G. 1. Position patient with medic's dominant hand on patient head, non-dominant had toward patient's feet.	10 secs	VISUAL: Orientation of patient relative to medic	Relative position of patient and medic	Failure to properly assume position
G. 2. Confirm location, accessibility of needed equipment in medic bag.	5 secs	VISUAL: Equipment locations	Necessary Equipment for Procedure	Failure to Confirm Equipment location
G. 3. Confirm that balloon holds air with syringe.	10 secs	VISUAL: Balloon inflation	Expected Balloon Shape when Filled	Misinterpreting balloon shape
G. 4. Visually ensure balloon is flat (not inflated).	2 secs	VISUAL: Balloon structure	Expected Balloon Shape When Empty	Misinterpreting balloon shape
H. 1. Perform laryngeal handshake using thumb, index, and middle fingers of non-dominant hand.	10 secs*	TACTILE: Thyroid rings, inferior border of the thyroid cartilage, superior border of the cricoid cartilage	Neck Anatomy	Grasping wrong area; grasping with wrong fingers
H. 2. Feel for structure using index finger of non-dominant hand.		TACTILE: Cric membrane depression, groove below notch	Neck Anatomy	Using incorrect finger; missing landmarks; interpreting false landmarks; feeling too high/low on trachea
H. 3. Identify hyoid bone – thyroid cartilage using index finger of non-dominant hand.		TACTILE: bumps and depressions	Position, structure of H-T cartilage	Identifying incorrect structure
H. 4. Identify cricothyroid cartilage using index finger of non-dominant hand.		TACTILE: bumps and depressions	Position, structure of cricoid cartilage	Identifying incorrect structure

H. 5. Identify cricothyroid membrane using index finger of non-dominant hand.		TACTILE: bumps and depressions	Position, structure of cricoid membrane	Identifying incorrect structure
I. 1. If possible, thoroughly clean tracheal area	5 secs	VISUAL: cleanliness of skin area	Neck Anatomy	Failure to clean area
I. 2. Mark the neck where you intend to cut – 2 finger breadth vertical line.	3 secs	VISUAL: adam's apple location; orientation of neck	Optimal line length	Line too short or long; position incorrect
J. 1. Use the Cric Knife with dominant hand to incise a vertical incision from mid thyroid cartilage to cricoid cartilage - about 2 finger breaths.	15 secs	VISUAL: skin opening, layers, underlying structures; TACTILE: underlying structures, scalpel blade, blade depth	Target end points for incision	Cutting too shallow or too deep; incorrect scalpel grip; using tip of scalpel; cutting too high or low; thyrohyoid membrane incision; cutting anterior jugular veins; fracture of cricoid or thyroid cartilage; excessive bleeding; tracheal injury
J. 2. Use Cric Knife to retract tissue layers	5 secs	VISUAL: skin opening, layers, underlying structures; TACTILE: underlying structures, scalpel blade, blade depth	Location and structure of tissue layers	Failure to retract tissue layers, failure to recognize anatomical landmarks
J. 3. Turn the Cric-Knife to a horizontal position over the cricothyroid membrane.	2 secs	VISUAL: relative position of knife and cric membrane	Location of cric membrane	Effecting excessive or insufficient rotation
J. 4. Push the blade downward with dominant hand, perpendicular to the trachea, until the blade is fully inserted, the airway is entered.	5 secs	TACTILE: poking through cric membrane, bone resistance	Location plane of trachea	Cutting too deeply, cutting too wide, severing blood vessels or vocal cords

J. 5. While maintaining downward force, slide the tracheal hook down the handle with your thumb until the hook is felt to enter the trachea.	6 secs	TACTILE: resistance change	Tracheal hook structure; desired resistance change	Losing control of incision; tearing ET tube balloon
J. 6. Confirm placement by moving the device along anterior wall of trachea to feel for the tracheal rings.	5 secs	TACTILE: bumps of tracheal rings	Area where tracheal rings exist	Failure to contact anterior wall; failure to listen for moving air
J. 7. Disengage the hook from the handle with non-dominant hand.	3 secs	VISUAL: relative position of components. TACTILE: cues of both components	Connecting mechanism	Dropping either component; failure to separate components
J. 8. Grab the tracheal hook with the non-dominant hand, lifting up on the thyroid cartilage.	7 secs	TACTILE: resistance when lifting cartilage	Flexibility of thyroid cartilage	Losing control of tracheal hook; twisting hook; entering subcutaneous layer
J. 9. Insert Cric-Key through incision with dominant hand.	5 secs	VISUAL: incision hole, tenting of skin; TACTILE: resistance of Cric-Key tube; KINESTHETIC: hand position	Proper orientation of Cric-Key	Tearing incision
J. 10. Once confirmed, advance Cric – Key to the flange.	2 secs	VISUAL: flange point	Recognition of flange structure	Insufficient insertion
J. 11. When placement is confirmed, use dominant hand to turn tracheal hook toward patient's shoulder and remove it.	3 secs	VISUAL: orientation of hook. AUDITORY: sound of rushing air to confirm correct placement	knowledge of desirable hook position	Failure to remove hook; tearing incision
J. 12. Remove bougie by pulling with dominant hand while securing Cri-Key with non-dominant hand.	3 secs	VISUAL: disconnection of bougie from Cri-Key; TACTILE: resistance when pulling bougie	Location of bougie knob	Failure to completely remove bougie
J. 13. Stabilize the Cric-Key.	20 secs	TACTILE: stability of Cric-Key	Procedure for stabilizing Cric-Key	Cric-Key too loose

J. 14. Inflate the cuff until resistance is met.	4 secs	VISUAL: connection of syringe and cuff, proper cc line, confirmation of removal; TACTILE: confirmation of removal	Inflation procedure	Failure to make connection; leaving syringe attached; allowing leakage or deflation
J. 15. Ensure bilateral chest rise.	5 secs	VISUAL: observing rise and fall of chest, misting in tube; TACTILE: proper attachment	Knowledge of respiration rate	Misinterpreting chest rise
J. 16. Auscultate lung sounds.	5 secs	AUDITORY: chest sounds	Signals for clear and congested lungs	Failure to detect sounds
TOTAL TIME FOR NOVICE PERFORMER	130 secs			

* Time estimate includes subtasks labeled H1-H5

References

Cannon-Bowers, J., Bowers, C., Stout, R., Ricci, K., & Hildabrand, A. (2013). Using cognitive task analysis to develop simulation-based training for medical tasks. *Military Medicine, 178*, 15–21.

Demirel, D., Butler, K. L., Halic, T., Sankaranarayanan, G., Spindler, D., Cao, C., Petrusa, E., Molina, M., Jones, D. B., and De, S. (2016). A hierarchical task analysis of cricothyroidotomy procedure for a virtual airway skills trainer simulator. *The American Journal of Surgery. 212*(3), 475-484.

APPENDIX B - Literature Review

INTRODUCTION

The use of simulation technologies for representing military medical tasks has burgeoned over the last five decades (1970 to 2022). As sensor sophistication, physical material manufacturing methods, and immersive technologies have improved and matured, simulation developers have explored a variety of approaches for medical simulation design. Many factors have driven such progress. Because of the cost and logistical challenges surrounding real patient training, the medical community at large has embraced the use of simulators for training critical tasks. Such simulators are particularly important for training trauma care tasks that demand precise execution while under stress.

Trauma care tasks of greatest concern relate to controlling massive hemorrhage, alleviating airway restrictions, managing respiratory distress, ensuring circulatory function, reducing hypothermia, treating head injuries, mitigating pain, administering antibiotics, caring for wounds, and applying splints. Airway management by surgical cricothyroidotomy is of particular concern because of the time demands for interdiction and the relative infrequency of training offered to medical personnel.

The following sections of this report are intended to provide a comprehensive review of recent literature (last 15 years) concerning simulation of surgical cricothyroidotomy. Articles reviewed concern the technological development of cricothyroidotomy simulation options, results of scientific comparison and evaluation of simulation options, and determination of the relative physical and functional fidelity represented by training simulators. Specific sections of the review address, in order, the construct of simulator fidelity, the history and practice of simulation-facilitated task training, comparison of simulators used for cricothyroidotomy training, the varieties of technologies currently used for cricothyroidotomy simulation, and the importance of assessing the degree of agreement between trainee competence and confidence during training. The report concludes by listing recommendations for administrators responsible for procuring simulation technology as well as training designers responsible for integrating cricothyroidotomy simulation devices within a training curriculum that includes didactic and hands-on experience.

Historically, tools have been generated to facilitate the simulator procurement process. For example, Ruisi, Wier, and Ziegenfuss (2017) documented the development of the Simulation Procurement Equipment Requirements Matrix (SimPERM). This tool allows procurement decision makers to specify the capabilities that are desired within a simulation platform. SimPERM then maps each requirement to existing simulation devices so procurement officers can make informed decisions about devices based on their functionality. Use of the platform has saved millions of dollars and has enabled simulator procurement decisions to be made independent of manufacturer inputs. Similarly, the Simulator Value Index (SVI) has been described by Rooney et al. (2018) as a tool for streamlining procurement decisions on the basis of 17 factors such as simulator cost and stability. Likewise, the SVI was shown to standardize and improve decision making for simulation professionals. Though both tools are clearly useful for procurement professionals, neither focuses on the realism or fidelity of simulation devices for representing the target task or procedure.

The literature review that follows is one of several tools one might use to establish a simulator's fidelity with respect to that which it is simulating. Literature reviews allow one to understand the way prior researchers have defined high-fidelity simulation with respect to a task as well as the landscape of available simulators and training methods.

Within the following paragraphs, we offer the results of a literature review that focuses on surgical cricothyroidotomy, a representative critical procedure of last resort for which multiple simulators exist. Our review begins with the construct of fidelity as it is referenced in the literature. From there, consideration is given to simulation training, varieties of simulation technology currently available, simulator comparison studies, and considerations of confidence and competence. We conclude the document with a list of best practices and recommendations targeted to simulation professionals and procurement administrators.

FIDELITY

The term fidelity has no single, widely accepted definition. In the domain of medical task simulation, fidelity is typically defined as how closely a simulator recreates cues present in a real-life patient scenario. This definition encompasses psychological, physical, and conceptual aspects of a simulator, as well as its ability to facilitate all or some responses from a trainee that could or should occur in reality. For a simulated training scenario to be considered "high fidelity," it would recreate many of the cues present in the true environment. The highest fidelity simulator would represent all the task cues, but such an exhaustive treatment is neither time nor cost efficient. As an alternative, the use of a cognitive task analysis (CTA) can help break down the cues for a task, as it can help identify the cues required for the appropriate decision to be made (Cannon-Bowers, 2013). CTA may also help determine at what level of fidelity each cue needs to be presented to produce effective task learning. This is similar to the method of fidelity needs analysis (Roza, 2004).

A study by Walsh, Heiner, Kang, Hile, and Deering (2013) provides a good application of high-fidelity sensory simulation. In a combat scenario necessitating cricothyroidotomy, participants saw flickering of lights to simulate explosions, smelled burned flesh-scented smoke in the room, and heard simulated gunfire and explosions (Walsh et al., 2013). The representation of such stimuli in the environment recreated the stress, dim lighting, loud noises and smells of an intense combat setting, where cricothyroidotomy often needs to be successfully performed, making it a high-fidelity simulation.

Fidelity is often broken down in the literature into three components: psychological, physical and conceptual factors (Kim, Park, & Shin, 2016). Physical fidelity includes the environment and equipment of the simulation, psychological fidelity relates to learner experience, motivation and emotion, and is the prompting of psychological responses mirroring the actual environment. Conceptual fidelity relates to the simulation providing the learner the opportunity to make proper connections between theoretical concepts and their meanings and relationships. In other words, the capacity of the simulation to make logical sense. For example, if a manikin is programmed to have a sudden rapid beating or fluttering of the heart, lack of breathing or pulse that simulates cardiac arrest, the simulation has high conceptual fidelity if it is interpreted by the learner as representing the concept of cardiac arrest (Kozlowski & DeShon, 2004; Paige & Morin, 2013).

There is argument about which aspect of fidelity deserves most focus when developing a simulation (Hamstra, Brydges, Hatala, Zendejas, & Cook, 2014; Kozlowski et al., 2004). Physical fidelity has been a focus historically, with the goal of selecting or building simulators that look and feel realistic. However, some researchers have called for a shift to an emphasis on psychological fidelity, accentuating function, and transfer of task elements. Instead of placing focus on feel and look, emphasizing whether the simulation evokes the central psychological concepts that would occur while performing the task.

Participant engagement is very important when thinking about medical simulation, and a consequence of increasing fidelity of simulation devices is the uncanny valley. It is the theory that if a robot or avatar is too similar to a human or reality, they evoke fear or discomfort instead of engagement. Moulage, the application of mock injury, is sometimes used to create realistic looking wounds and injuries for simulation. In a study by Stokes-Parish, Duvivier, and Jolly (2019), experts in moulage ranked its' authenticity, and what they considered most important: anatomical correctness, detail, position, and color of the moulage (Stokes-Parish et al., 2019). However, they also discuss that if a moulage is too realistic or disturbing, it can take away from the fidelity of the simulation by deterring participant engagement.

Compared to low fidelity simulators, high fidelity simulators produce better performance and are preferred by participants (Johnson, 2017). However, this finding likely interacts with expertise, because experts are more likely have a difficult time transferring their skills from the simulator to the real task if the simulator is highly unrealistic. A simulator with more faithful appearance, feel or anatomy is rated higher for usefulness. Research findings show that regardless of experience, overall groups perform best on a high-fidelity simulator (Cho et al., 2008; Johnson, 2017).

For cricothyroidotomy simulation, a popular choice is to use animal models, due to likeness with human anatomy and skin feel. The most frequent choice is the use of porcine tracheas, but other animals (e.g. chicken skin, goat larynx and trachea) have also been supported (Senuren et al., 2020). Porcine models, often consisting of skin, larynx, or tracheas, are generally considered high fidelity, though studies have reported participant ratings that indicate less realism than anticipated (Katayama et al., 2019). There is, however, much data backing their use. Participants commonly report porcine skin as realistic when compared to human skin (Cho et al., 2009). Anesthetized animals even provide aspects of the whole surgical experience: bleeding, breathing to show success of intubation, and opportunity to monitor vital signs (Heard et al., 2020). However, there is variation; some authors have characterized porcine tissue as too thick or tough to properly train a participant to apply the correct amount of force to cut through the skin. Also, animal models are not standardized in size and shape as synthetic models are (Mabry, Frankfurt, Kharod, & Butler, 2015). Last, even if the skin resembles human tissue, porcine trainers lack a considerable number of components necessary to replicate the cricothyroidotomy procedure (e.g., face, lungs, etc.).

There is also the option to augment animal parts, such as the previously mentioned chicken or porcine skin. Kei and colleagues (2019) created a breathing and bleeding, high fidelity model for cricothyroidotomy, using pork belly skin. It was considered high fidelity because the bleeding led to a more stressful, more obscured, and therefore realistic environment during the training. The authors' approach also lowered the trainer cost because they used only a small part of the animal in conjunction with other readily available items (e.g., a plastic endotracheal tube). Participants in their study described the pork belly skin as very realistic.

As noted above, arguments have been made against animal models. These include claims that they are not able to properly simulate human emergency conditions or the human patient scenario, and a call for a shift away from animal models, especially live, for ethical and cost reasons. In addition to animal models, researchers commonly use human patient simulators.

Widely used for a variety of medical task training, human patient simulators are manikins of the entire human body. They vary greatly in look and function. Those used in military or emergency medical training may have wounds, missing limbs, or severe facial injury, increasing situational fidelity. Also, some manikins have the capability to simulate breathing that might indicate to participants when and airway is established. Others are capable of bleeding, which obscures the work area and would occur in a real medical emergency context. All the previously listed cues contribute to a higher fidelity simulation. They provide the ability to train without the risk of operating on a patient, or the ethical issue or repeated cost of animal use. A study by Boet and colleagues (2011) found a single session on a high-fidelity simulator was enough to improve skills and aid in skill retention for a least a year.

As reusable manikin and virtual reality simulation technology continues to evolve, authors such as Gala and Crandall (2018) suggest shifting to Human Patient Simulators (HPS). These manikins have anatomy closer to humans than animal models. They also offer capabilities like haptic feedback and can be tailored to represent specific patient demographics, maladies, and medical situations. Also, the authors note studies that have shown HPS manikins to be equally or more effective than animal models during an Advance Trauma Life support skill training.

Problems associated with reliance on HPS are their high cost, and limited reusability. Certain parts, once operated on, must be replaced. Though these trainers are rated higher when compared to more simple trainers on fidelity, there are often complaints about realism of the skin and tissue texture. Participants often rate higher realism for the skin of porcine trainers as opposed to the manufactured skin used on manikins (Cho et al., 2008). Also, some HPS trainers lack the true anatomy of a human. One study by Schebesta and colleagues (2012) compared tomography scans of the neck anatomy of several HPS and compared that to human scans, finding significant differences. Diameter of the tongue, oral airspace, and the distance from the epiglottis to the posterior pharyngeal wall were different in all manikins when compared with the actual patients.

When the fidelity levels of HPS are compared to animal models, animals are often rated higher on realism or fidelity. This is mainly because though medical simulation companies aim to mimic the tissue and other characteristics of a human patient, it is not the same as the organic qualities an animal has. A study by Cho and colleagues (2009) compared a porcine model consisting of a porcine trachea, larynx and skin, to a manikin that had a synthetic rubber trachea and skin. Participants in the study rated realism and preference higher for the porcine, considering it closer to reality and to anatomy and skin feel.

Clearly, some schoolhouses (particularly those with connections to medical schools) rely on cadavers for high fidelity training. Cadavers feature high fidelity due to their human anatomy, accurate landmark identification and patient variability (Takayesu et al., 2017). However, they lack anatomical responses generated by live patients, and embalmed cadaver tissue texture differs from that of a live patient. Cadavers are also expensive, costing several thousand dollars each. Importantly, they are also single-use trainers unless they are supplemented with synthetic technology. Toward that end, synthetic human cadavers exist, and their tissue is reusable, repairable, and does not degrade. This solves many issues; however, the cost of synthetic cadavers is prohibitive for many training centers. For example, one SynDaver synthetic cadaver costs approximately $70,000.

Low fidelity models, some of them referred to as ‘homemade’ models, also have their place in cricothyroidotomy training. Though they are more cost-effective than other methods, they often represent very little of the correct anatomy. Homemade simulators sometimes incorporate a combination of low- and high-fidelity components. For example, Kei and colleagues (2019) created a breathing and bleeding cricothyroidotomy model using pork belly, red saline, a bag valve mask and endotracheal tube. This model offered high fidelity experiences to participants, bleeding obscuring the area, and active respiration to reflect whether an attempt was successful. They found this model produced a stressful environment for participants, who described the feel of skin and palpation for landmarks as realistic. Generally, homemade models score lower on realism, but may perform close to or as well as commercially available options for certain procedure aspects. This suggests that complete anatomical functionality and accuracy may not always be necessary to successfully learn skills (Varaday et al., 2004). Bryant and colleagues (2017) have noted that inexpensive, homemade options (often costing less than $75) constitute a reasonable solution for repetitive practice during skill acquisition. Friedman and colleagues (2008) compared cricothyroidotomy training on high and low fidelity ($10) models and found the same effect on rated skill acquisition from both.

One pervasive challenge is that perceptions of fidelity often vary across users. A simulator that is rated positively by one person may be rated negatively by another. Shefrin and colleagues (2015) found great variation in fidelity ratings for homemade models. Curiously, the fidelity of some simulators may be considered high during training, but ineffective for skill transfer.

New developments in technology have also led to the ability to develop high fidelity simulation but in different ways, including at a lower cost. Three-dimensional printed models have begun to be adopted more widely for medical training, as well as simulation technologies that rely on immersive technology (augmented and/or virtual reality). Many of these constitute high fidelity, but low-cost alternatives to a full or part body manikin trainer. Three-dimensional, printed trainers include many abilities manikin trainers have, including the potential to bleed when cut, accurate anatomical landmarks, layers of bone and tissue, and support for task repetition, which is important for establishing skill competency (Han et al., 2019; Hughes et al., 2018; Shaw & Hughes, 2020). To mimic the feel of cutting into a human's neck, an area of discomfort for many in emergency medicine, Duggan et al. (2017) created a 3D printed model that attached to the neck. It provided a unique, high-fidelity experience and haptic feedback from a human patient.

Virtual reality and other digital media trainers can provide widespread training without the need for expert facilitators or physical equipment. VR and similar programs can recreate an actual performance environment in a way that still prompts psychological responses like the real environment (Kozlowski et al., 2004). As early as 2005, Lui and colleagues (2005) argued for use of computer-based simulations because they allow for customizable scenarios. Using modern technology, nearly any scenario or patient can be created and added to existing software. A virtual airway skills trainer created by Sankaranarayanan et al. (2022) provided a computer display of virtual anatomic models and tools for the procedure.

SIMULATION TRAINING

Many medical personnel do not feel confident in their skills, yet new developments in technology enable effective simulation programs that can be widely distributed. Methods for cricothyroidotomy simulation vary, but well-done scenarios often share common features: opportunities for feedback, repetitive practice, including lecture modules prior to applied practice, utilizing pre- and post- testing, and inducing a stressful scenario environment. Grading methods vary but the use of task specific checklists and global rating scales are frequently employed and have been validated by prior research and adopted by organizations such as the Tactical Combat Casualty Care (TCCC) committee.

Simulation based medical education (SBME) has become common for medical training. Technology has evolved, making it highly flexible and adaptable for many patient scenarios and experience levels. It is highly effective in skill acquisition (Petrosoniak et al., 2019). Simulation training can provide more frequent training to medical personnel, reducing accidents and increasing performance during infrequent emergency procedures like cricothyroidotomy. In many organizations, emergency medical staff do not train for cricothyroidotomy regularly and do not feel confident performing the procedure. A study surveying paramedics found 73% felt they were not adequately trained, and 40% felt they could not correctly perform the procedure (Carvey et al., 2020).

Common methods used for simulation training vary, but many are taught via a combination of didactic, instructor-led lecture and applied practice. A lecture or similar presentation of information prior to a simulation training session is beneficial. A study by Ghosh and Chaudhury (2021) provided a 2-hour integrated lecture on anatomy and procedure for cricothyroidotomy, then selected students to practice. Post-survey responses showed knowledge in anatomy and confidence performing had significantly increased.

Instructor facilitated simulation training sessions have been found to be valuable. The ability to ask questions and gain feedback is useful before, after, and during the simulation. Debriefing by instructors prompts reflection and facilitates more in-depth discussion that complements training with a simulation device (Nguyen et al., 2019). When comparing training groups with and without an instructor, Mikkelsen et al. (2008) determined the role of the teacher was critical. Instructors provided feedback, examples, demonstrations, immediate answers and clarification, and prompted questions. Group members who learned without an instructor found their sessions less helpful, challenging, and socially engaging. Furthermore, trainees found instructor-less sessions to be less challenging and informative.

Feedback to learners is essential during simulation training, whether from an instructor or another source. In-scenario instruction and feedback from facilitators can shape the process and change the trainee experience (Escher et al., 2017). Debriefing by instructors prompts reflection and discussion about learner reasoning and decision making (Nguyen et al., 2019). In some cases, the simulator itself can provide feedback. Netto et al. (2015) proposed an incision sensor measuring incision path for cricothyroidotomy that could provide feedback about the procedure, help trainees understand their mistakes, and provide further insight in a training scenario, all of which is vital in a learning environment.

A common alternative to instructor-led training is the use of video or modules, presented either in the context of an online simulation, or in conjunction with in-person training or a synchronous lecture. Distributed training with systems that allow geographically separated trainees and instructors expand learning opportunities to more participants (Kozlowski et al., 2004). This method of training has become more useful due to advancements in technology, (e.g. apps, media, virtual reality, internet video conferencing) and the recent COVID-influenced drive for distant learning. Presenting lesson material remotely may also save the time of senior instructors. Using a flipped classroom model, where video lectures or modules are used prior to meeting for a hands-on simulation training has also been found useful (Shefrin et al., 2015; Glass et al., 2019). In a study by Melchiors et al. (2016), simulation training using instructional video combined with hands-on training was successful in improving skill level and performance.

A common format among training programs is using pre- and post-testing to assess progress. The test is often a hands-on performance of cricothyroidotomy using a simulator after which scores are computed (Friedman et al., 2008; Gauger et al., 2018). Testing can be useful for instructors to design debriefing questions (e.g. What went well? What did you do to successfully intubate?). Such questions help evaluate the training program and break down what was learned by trainees (Murray & Konia, 2010). Delaying post-training testing can show retention of learned information. By doing so, studies were able to determine that the cricothyroidotomy skills learned were retained from one month, 120 days, and up to a year, without deterioration (Melchiors et al., 2016; Muller et al., 2020; Boet et al., 2011 respectively). In training performed using a virtual reality trainer or similar multimedia device, grading methods may be automated using equipment algorithms, or computed manually by reviewing hands-on performance. A study by Qi and colleagues (2021) used a VR simulator and gave participants a final trial on the immersive device that was considered the formal test of their ability. In another study utilizing VR, authors used a TraumaMan HPS to evaluate trainees' transfer of skills during a post-test (Sankaranarayanan et al., 2022).

There is also the option to utilize a longitudinal training approach, with multiple sessions over a longer span of time. Nguyen et al. (2019) evaluated a four-year curriculum in airway crisis management, with testing sessions separated by six to 12 months and increasing in complexity. Trainees participated in and observed simulation scenarios, including lecture and hands-on training. This was an alternative to a typical stand-alone session, and following all modules, participants reported significant improvement in airway management skills.

Another method proven to be useful during training is using multiple training sessions. Allowing trainees to practice until they are successful or confident has been shown to improve performance and transfer of skills (Sankaranarayanan et al., 2022; Scott-Herring et al., 2020; Qi et al., 2021). An important question surrounding the use of multiple practice sessions concerns when to expect stability of performance. Generally, authors have recommended the use of at least five sessions to establish such stability during cricothyroidotomy training (Scali & Espinosa, 2020; Shetty, Nayyar, Stachowski & Byth, 2013). Adherence to this guideline benefits muscle memory, cognitive skill acquisition and confidence (Bryant et al., 2017). In a review by Cook et al. (2013), the authors included task repetition in their list for best practices in effective simulation education.

When implementing a cricothyroidotomy simulation, it is common to begin scenarios with a "cannot intubate, cannot oxygenate (CICO)" condition, necessitating an immediate cricothyroidotomy attempt and catalyzing a response similar to a real CICO situation (Gauger et al., 2018; Paige et al., 2009). The increased stress of the situation is important to challenge comfort, learn stress management skills, and practice maintenance of communication, along with the ability to complete the procedure (Duggan et al., 2017). Murray and Konia (2010) provide guidelines to designing a difficult airway simulation scenario with a HPS.

Grading methods used in simulation education vary. Common measures are procedure time, success of performance, and successful intubation. For example, many researchers use both a task specific checklist and global rating scale (Schmitz et al., 2014; Boet et al., 2011; Friedman et al. 2008; Gauger et al. 2019; Siu et al., 2010; etc.). A global rating scale uses general descriptions and gauges overall procedural performance for a subject, and a task specific checklist goes through the steps in the task and gives a score of 0, 1 or 2 for each, 0 if it was not performed, 1 if performed poorly, 2 if performed well (Boet et al., 2011). Friedman and colleagues demonstrated face and content validity and inter-rater reliability for both the Global Rating Scale GRS (GRS; see Figure 1) and the Examiner's Checklist (EC; see Table 1) (Boet et al., 2011; Friedman et al., 2008).

Preparation for procedure	1	2	3	4	5
	Did not organize equipment well. Has to stop procedure frequently to prepare equipment		Equipment generally organized. Occasionally has to stop and prepare items		All equipment neatly organized prepared and ready for use
Respect for tissue	1 Frequently used unnecessary force on tissue or caused damage	2	3 Careful handling of tissue but occasionally caused inadvertent damage	4	5 Consistently handled tissues appropriately with minimal damage
Time and motion	1 Many unnecessary moves	2	3 Efficient time/ motion but some unnecessary moves	4	5 Clear economy of movement and maximum efficiency
Instrument handling	1 Repeatedly makes tentative or awkward moves with instruments	2	3 Competent use of instruments but occasionally appeared stiff or awkward	4	5 Fluid moves with instruments and no awkwardness
Flow of procedure	1 Frequently stopped procedure and seemed unsure of next move	2	3 Demonstrated some forward planning with reasonable progression of procedure	4	5 Obviously planned course of procedure with effortless flow from one move to the next
Knowledge of procedure	1 Deficient knowledge	2	3 Knew all important steps of procedure	4	5 Demonstrated familiarity with all aspects of procedure
Overall performance	1 Very poor	2	3 Competent	4	5 Clearly superior

Figure 1. Global Rating Scale for Cricothyroidotomy (taken from Friedman et al., 2008).

Table 1. Examiner's Checklist for Cricothyrotomy Performance (taken from Friedman et al., 2008).

1. Correctly identifies cricothyroid membrane

2. Stabilizes the cartilage.

3. Makes an incision in the midline using the #15 short handle scalpel blade.

4. Advances syringe attached to the introducer needle and catheter through the incision and cricothyroid membrane into the airway at a 45-degree angle.

5. Verifies entrance into the airway by aspiration syringe resulting in free air return.

6. Removes the needle, leaving the catheter in place.

7. Advances the soft, flexible end of the wire guide through the catheter and into the airway several centimeters.

8. Removes the catheter, while maintaining control over the wire guide and leaving it in place.

9. Advances the handled dilator, tapered end first, into the connector end of the airway catheter until the handle stops against the connector.

10. Advances the assembly over the wire guide until the proximal stiff end of the wire guide is completely through and visible at the handle end of the dilator.

11. Maintaining wire guide position, advances the assembly over the wire guide with a reciprocating motion, and completely into the trachea.

12. Removes the wire guide and dilator.

- 0 = did not perform.
- 1 = inadequately performed.
- 2 = adequately performed.

Some simulation training studies have used dissection and evaluation of cut and intubation length, damage to surrounding tissue and whether the membrane was punctured to determine success. Following performance, graders evaluate the neck tissue to assess cut placement, depth, accuracy (Carvey et al., 2020; Hall, 2011; Sankaranarayanan et al., 2022; Ghosh & Chaudhury, 2022; Umek et al., 2020). Ahn and colleagues (2019) proposed the creation and use of an incision sensor that can track the incision path for cricothyroidotomy to determine if it is correct and provide feedback. However, this is not a viable grading option across simulators, especially those that lack realistic tissue.

Multiple studies have reported results affected by age and experience level of trainees. A cross-sectional study by Siu et al. (2010) found that age affects learning and performance of cricothyroidotomy, with age defined as groups older and younger than 45 years. Training was standardized; however, increased age of participants and years from residency were associated with decreased proficiency to perform the procedure. Another study by Qi et al. (2021) compared more experienced to less experienced trainees, finding that those who had performed at least five cricothyroidotomy procedures performed significantly better than those who had performed fewer than five. Such results suggest that number of simulated training sessions should be increased for older or less experienced trainees.

A final point regarding simulation-based education training is the use of mastery learning and deliberate practice. Deliberate practice is an approach that utilizes engagement in activities to improve performance through interactive practice (see Figure 2). Incorporating feedback and refinement of skills, deliberate practice is purposeful and systematic (Petrosoniak et al., 2019; Asselin et al., 2021). Mastery learning involves breaking down broader tasks into smaller 'micro-skills' that are more complex and allow for closer practice and study. Though implementation of such approaches needs further investigation, early results suggest that designing curricula to reflect both approaches (whether together or separate) yield positive results for skill mastery. However, doing so may be a complex and time-consuming undertaking (Petrosoniak et al., 2019; Asselin et al., 2021; Issa et al., 2021).

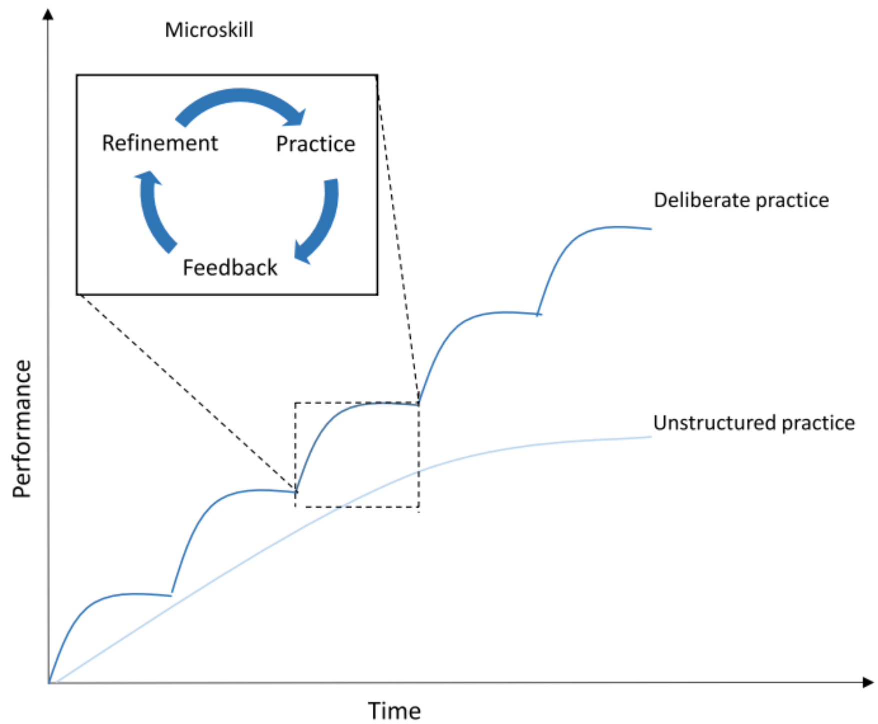


Fig. 1 Deliberate practice model. Deliberate practice occurs over time with repeated feedback and refinement of microskills, combined together resulting in improved task completion

Figure 2. The Deliberate Practice Model (taken from Petrosoniak et al., 2019).

VARIETIES OF SIMULATION TECHNOLOGY

A variety of alternative simulator technologies have been introduced, reflecting technological advancements and the high cost and low reusability of many traditional simulators. Though traditional models like cadavers, animal models, and manikins are still commonly used, advanced technologies such as virtual/augmented reality and 3D printed anatomical parts have gradually been adopted as well. In certain cases, trainers have also modified conventional animal models and artificial manikin simulators to increase their utility and applicability for certain procedures.

The use of cadavers has been a preferred method of training for decades. Cadavers may represent a variety of patient demographics with complete anatomical correctness, though they may lack the feel of real living tissue and physiological responses of live human patients (e.g., haptic feedback, bleeding). Cadavers are often used and preferred as a standard for testing. However, the logistics and cost of acquiring, using, and disposing of cadaver tissue renders them a less viable choice for many programs.

Regarding animal models, trainers and technology procurement officers face the choice of using an entire model or isolated components like tracheas, targeted organs, or outer skin layers. In recent years the introduction of pieces of animal models into other trainers has proved useful. For example, some trainers have adapted the use of chicken skin to serve as a more human-feeling tissue rather than the rubber or plastic skin of many manikin models (Senuren et al., 2020). Others have implemented a porcine trachea into an opening of a 3D printed model for a more realistic, and affordable, airway simulator (Issa et al., 2021). Netto and colleagues (2015) took a porcine trachea and added a plastic bag to the end and covered it with a surgical cloth to simulate chest movement. A rubber glove was placed internally to represent a 'new' cricothyroid membrane after one procedure so the model could be reused, and the model was covered with porcine skin for a more realistic feel. This approach increased the realism of the porcine trachea and allowed the model to be reused. Combining cheaper materials, like a Styrofoam head, and implementing animal skin or trachea may result in an affordable model that is often reusable, though perceived realism may suffer.

Artificial manikins have gained abilities over time. The most frequently cited manikin in the literature reviewed was the Laerdal SimMan, followed by the Simulab TraumaMan. The SimMan and manikins like it can display a variety of symptoms, support use of moulage to depict injuries, simulate respiration, and allow for haptic feedback. The most current SimMan can secrete sweat and tears, spontaneously breathe, allow dilation of pupils, and reflect many other physiological and physical states. Such manikins also coordinate with online software platforms that allow diverse clinical scenarios to be represented. The Laerdal scenario cloud provides hundreds of 'expert-validated' scenarios ready to use with a SimMan. Choice of scenarios and functions may result in perceptions of high or low manikin fidelity (John et al., 2007).

A derivation of rubber or plastic manikin models are simulators that feature artificial flesh. For example, the SynDaver simulators are designed to have realistic tissue, created with water, salt and fibers meant to mimic as closely as possible the makeup of live human tissue, and mechanical and physical properties (https://syndaver.com/product/syntissue-surgical-model/). Each individual organ's tissue is created and refined; organs are removable, replaceable and in some cases reusable.

3-D Printed Models. An increasingly popular method for making more affordable, yet still realistic simulators, is to implement 3D printed components. 3D printed tracheas are often used for cricothyroidotomy. Multiple studies utilize a free, online available 3D printing model for a trachea (Calvo et al., 2021; Kei et al., 2019; Huang, 2021). That model was initially provided by Duggan et al. (2017); the online file is free of charge at the airway collaboration website (http://www.airwaycollaboration.org/3d-cric-trainer-1). Other researchers have created and utilized their own models; Katayama et al. (2019) created a novel 3D printed trachea based on tomography scans. Other materials used in conjunction with the 3D printed models have included animal meat or skin, layered silicone pads (e.g., Dragon Skin, EcoFlex), self-adhering medical wrap (Coban), and Durapore surgical tape. Many solutions incorporate silicone because it resembles the texture and elasticity of actual human skin.

Models can be partially or entirely made of 3D printed elements, and 3D printed models may possess high-fidelity abilities, like bleeding and breathing. One solution evaluated by Shaw and Hughes (2020) includes 3D printed housing, 3D printed anatomic trachea, and layered silicone tissue, with bleeding ability. In some cases, artificial blood can be inserted into the layered tissue and will release when cut (Calvo et al., 2021). Using a bag valve mask and endotracheal tube, Kei et al. (2018) created a 3D printed trachea that also expelled air when punctured. Some model components are reusable, with only the ‘skin’ having to be replaced, Shaw and Hughes (2020) produced a model that allowed each tissue section to be used 3 times before being discarded. Another benefit of 3D printed models is that they can effectively represent special anatomical cases or conditions. 3D printing has been used to create models for cases like congenital heart defects and other variations from normal anatomy. For cricothyroidotomy, using thicker ‘skin’ material, whether that be meat, silicone, or medical wrap, can represent an obese patient, where palpation is more difficult, and incisions must be made more deeply (Kei et al., 2019; Duggan et al., 2017). For entirely 3D printed models, more layers could be used for a similar outcome.

Perhaps the greatest benefits associated with 3D printed models are low cost and improved accessibility. Models range in cost from around $20 to $70, depending on the detail of the model, size, and materials used, but they are still considerably cheaper than many commercial trainers. They also may be assembled quickly after printing is completed. A study by Gauger et al. (2019) implemented 3D printed, high-fidelity cricothyroidotomy trainers in Ethiopia, where there is a shortage of trained health professionals, equipment and supplies. The authors implemented a low-cost 3D printed simulator using a novel laryngotracheal model, created using medical grade silicone. Statistical results from the training program suggested improved completion time post-training, as well as trainee confidence. Post-training questionnaire results showed that 3D printed models were highly rated, even in comparison to conventional trainers (Katayama et al., 2019). The 3D printed models were found to have increased user comfort as well (Takayesu et al., 2017; Shaw & Hughes, 2020; Gauger et al., 2019). They have also been found comparable to animal models (Huang et al., 2021). Generally, 3D printed models are easy to create, use affordable materials, and can provide a trainer solution that is broadly accessible.

Simulators Featuring Immersive Technology (AR/VR/XR/MR). Mixed, virtual, augmented, and extended reality technology have become popular for medical simulation training, and cricothyroidotomy is no exception. Immersive simulator solutions can replicate many of the capabilities of traditional animal and commercial models. They are particularly useful for medical simulation because they allow virtual information to be overlaid onto a physical manikin. Alternatively, solutions may be completely immersive, featuring an entirely virtual surgical experience. Immersive solutions may be used for guidance, team practice, communication, and to facilitate repetitive practice without the need for new materials. As technology continues to improve, the fidelity of visual experiences continues to increase.

Training utilizing immersive technologies has often relied on the use of head mounted displays in the form of glasses or goggles. In training scenarios where one medical professional guides a learner through a procedure, the glasses allow the remote expert to see a broadcast of what the participant sees as he or she performs the procedure. This is useful for feedback as well as for providing support and training to remote trainees. In a study by Miller et al. (2018), when guiding a physician’s assistant through a combat casualty care injury, an orthopedic surgeon was able to superimpose visual and aural instructions to the learner. Rojas-Munoz et al. (2020) evaluated an augmented reality system with a head mounted display that allowed a remote expert to guide first responders through a cricothyroidotomy scenario. The expert was able to create instructions for the procedure, and those instructions projected onto the operating field for the trainee to see. Immersive technology offers the ability to overlay information onto a trainer or manikin, including procedural directions, patient vital signs, inventories of available tools, and temporal status.

As previously stated, some immersive simulators have similar capabilities to other high-fidelity, commercial simulators. Designers have enabled bleeding, breathing, haptic feedback, and real-time procedural feedback and grading. Lui et al. (2005) described an early computer model that allowed the participant to see a virtual patient and interact with a menu to palpate, incise, widen the incision, and intubate. The extent of bleeding varied depending on the depth of the cut. Comparable technology evaluated more recently by Sankaranarayanan et al. (2022) used force sensors under a palpation interface that allowed users to hold a haptic device and choose tools to compete cricothyroidotomy. The system then automatically recorded performance metrics for assessment. Another model with VR capability was the CricSim VR system, which used a virtual patient for cricothyroidotomy training. The system showed a 3D image of a patient's head, torso, and neck. The patient was presented with a gunshot wound and CICO conditions. CricSim operated with stereoscopic 3D imaging through use of stereographic shutter glasses and haptic VR provided by two haptic interface devices (Proctor & Campbell-Wynn, 2014).

The use of serious gaming technology has also been adopted to provide real-world like experiences to individuals and teams for simulation, education and training (Bowyer et al., 2008). Because procedures such as surgical cricothyroidotomy are often accomplished by teams of medical professionals, practicing team communication through a video game environment is useful. BreakAway Games is a company that applies gaming software to simulation, and they are active in both military and healthcare training. Their solutions simulate navigating an operating environment, practicing specific medical procedures, and interacting with a variety of patients. Technology also exists for entire virtual environments, turning a room into a theater with the illusion of immersion by projection of images and use of VR glasses. Using gloves and handheld equipment, a user can interact with virtual objects. This could be an operating room with a virtual patient, or a military medical center, with added sound effects and smells, like the sound of gunfire or smoke to increase immersive feel and fidelity (Bowyer et al., 2008).

The use of wearable sensors can augment a learning experience through physical and virtual means. Using magnetic resonance imaging, Shen et al. (2016) created a prototype for a cricothyroidotomy simulator that uses virtual reality. Computerized mesh models were created with the use of the MRI images and then used as virtual airway models, whose 3D printed parts were integrated with sensors to track force and stress on the simulator, tissue deformation and movement. The authors suggested adopting mixed reality glasses to overlay multimodal information into one view, like other simulators previously mentioned. The embedded sensors can make models capable of objectively measuring trainee performance during practice. The solution also included electromagnetic sensors that measure the path of an endotracheal tube as it passes through the airway. This could potentially broaden the availability of scoring methods. Immersive simulation technology continues to develop, with the promise that solutions may represent viable alternatives to traditional simulation models.

Unique Models. Some recent cricothyroidotomy simulation solutions and technologies are unique in their design and materials. Notable in this category is a cricothyroidotomy model that is made of edible components, including a tortilla, peanut butter, gummy worms and other candy (Bryant et al., 2017). The model costs less than $3 to build. Though performance times and rankings for this model were degraded relative to porcine or plastic models, acceptance was indicated by participants. The authors concluded that development of muscle memory and skill acquisition that comes from repeated practice of the procedure may be feasible with such cheap and easy-to-build models. Other models made of affordable materials have also been described, including one that includes plastic breathing tubing, sticky tape, gauze, and the inner parts of rolls of tape to represent tracheal and cricoid rings (Varaday et al., 2004). The possibilities for creating a homemade trainer are numerous, as is evident by these two models.

A cricothyroidotomy simulator can be made using a wide range of materials, from candy to medical equipment. Variations of 'homemade' models have been described in the literature, often to save money and provide more opportunities for practice. Technology for cricothyroidotomy simulation is also continuously developing, from 3D printing and mixed reality to sensors capable of grading, looking at incisions, and providing feedback on palpations.

SIMULATION COMPARISON

Varying methods of simulator comparison are useful for demonstrating superiority under a variety of circumstances and include a range of approaches. Studies implementing these methods also show that many simulators for cricothyroidotomy training are comparable, and not dependent on high-level fidelity or complexity. The main difference pertains to a lack of realism for less complex trainers and manikins compared to animal or cadaver models, and likely reflects novelty or familiarity. Results suggest physical or technology differences across cricothyroidotomy simulators equate to negligible differences in performance or subjective ratings.

To compare simulators for medical training, investigators have used a range of measures. The most common metrics reported in the literature are success of procedure or correct tool placement (John et al., 2007; Katayama et al., 2019; Pandian et al., 2011; 2021; Quick et al., 2014; Umek et al., 2020), procedure time (Friedman et al., 2008; Hall, 2011; Huang et al., 2021; John et al., 2007; Katayama et al., 2019; Pandian et al., 2020), and simulator realism (Calvo et al., 2021; Cho et al., 2008; Hall et al., 2014; Hughes et al., 2018; Katayama et al., 2019; Takayesu et al., 2017; Varaday et al., 2004). Though employed less frequently, other useful metrics include learner preference (Hall et al., 2014; Cho et al., 2008; Walsh et al., 2013), difficulty with skin penetration (Cho et al., 2008; Cook et al., 2007), ease of landmark recognition (Cho et al., 2009; Qi et al., 2021; Takayesu et al., 2017) and confidence of trainees (Hall et al., 2014; Katayama et al., 2019; Takayesu et al., 2017). In most cases, investigators have used multiple measures to compare simulators. This practice is important given the potential for metrics to capture variance independently or to overlap.

Despite variation from study to study, success of placement is determined by evidence of obtaining an airway (evident through respiration) or penetration of the cricothyroid membrane. A failed cricothyroidotomy would occur if the incision was not completely through the membrane, the wrong site was penetrated, or if insertion is not truly in the trachea or follows an incorrect (Katayama et al., 2019). Procedure time is often measured from when trainees are told to begin or display motion indicating that they are beginning (e.g., lifting their hands, picking up tools). Participants are timed until success, or until they believe they have completed the procedure. It is frequently the case that there would be a time limit in which trainees must complete the procedure, or it would be deemed a 'failure', since cricothyroidotomy is a time-critical procedure. The construct realism (fidelity) is among the more varied across studies and is often dependent on features of the simulator. For example, ratings of realism may reflect presence of real or artificial blood (which frequently obscures the workspace). Other ratings may reflect general realism of a simulator. Rating scores are usually done via a Likert-style, visual analog, or open-ended questionnaire.

Scores from rubrics are also used to compare participant progress made on simulators. The Objective Structured Assessment of Technical Skills (OSATS) is commonly used. The validated OSATS includes 10 questions (Pandian et al., 2020). Nine rate each procedure step as unsatisfactory or satisfactory, one question is a global assessment of the participants' performance by rating it overall on a scale of 0–5 (0 = failure, 5 = exceptional). Also useful in comparison are the GRS (discussed previously) and the Tactical Combat Casualty Care (TCCC) rubric. The validated TCCC rubric has 13 items that assess the steps of cricothyroidotomy and result in ratings of "unsatisfactory" or "satisfactory" for each step (Pandian et al., 2020). Likert scales are most frequently reported as rating and comparing instruments (Calvo et al., 2021; Hall et al., 2014; Varaday et al., 2004; Paige et al., 2009; Shaw & Hughes, 2020). Visual analog scales are used less frequently (Takayesu et al., 2017; Hughes et al., 2018, Cho et al., 2008).

Simulator comparisons between low-fidelity, simple models and commercial manikins, cadavers or animal models have been widely reported in the literature. Early comparisons of homemade simulators and commercial manikins revealed markedly similar scores and had little if any significant differences (Friedman et al., 2008; Varaday et al., 2004). The main differences concerned realism. Results suggested that homemade, low fidelity models are useful and that full anatomical correctness may not be necessary for adequate training. A study comparing performance on cadavers versus a low-fidelity model found that high performance on the model correlated with high performance on the cadaver (Melchiors et al., 2014). In a study by Jing et al. (2021), researchers compared a novel, synthetic, bleeding model to a porcine tissue training model, finding similar scores for realism and improvement, but trainee preference for the porcine model.

Comparing commercial manikin trainers (or HPS) to cadavers, learners prefer cadavers for their higher tissue fidelity and anatomic landmarks, and for increasing their "comfort" (a construct that reflects contextual realism; Takayesu et al., 2017). When comparing manikin trainers to animal models, differences between the groups were small or non-significant, but animal models were preferred and tend to score higher in realism (Pandian et al., Cho et al., 2008; Hall et al., 2014; Hall, 2011).

For 3D printed model comparisons, Katayama et al. (2019) compared a 3D printed model (incorporating elements of a pig trachea) to conventional, commercial simulators, finding no significant differences in success rate or time. A similar study compared a 3D printed model to an animal model (porcine trachea) and found no significant differences in success rate or time (Huang et al. 2021). Finally, comparing a bleeding 3D printed model to non-bleeding models previously used by participants, raters indicated that the bleeding model was preferred to non-bleeding cadavers, porcine tracheas, or manikin models (Hughes et al., 2018).

VR and AR model comparisons are difficult to find given the novelty of such technology. However, as technology develops there may be wider adoption of augmented overlay technology onto medical manikins. Sankraranarayanan et al. (2022) recently compared a group trained with a VR system to a control group that experienced no training. All members of the immersive simulation group reached proficiency through use of the VR cricothyroidotomy training system. More research is needed to compare VR methods of training to more frequently used or commercial methods to determine technology preference and efficacy.

CONFIDENCE AND COMPETENCE

Because the consequences of poor performance are dire, it is appropriate to differentiate learner ability to perform surgical cricothyroidotomy (competence) from the learner's perception of procedural self-efficacy (confidence). Indeed, prior research has explicitly examined predictors of one's willingness to initiate cricothyroidotomy, examining the effectiveness of training programs to improve one's confidence (Berwick et al., 2019; Gauger et al., 2018; Ghosh & Chadhury, 2021; Scott-Herring et al., 2020; Zhang et al., 2021). A surprisingly low percentage (55%) of surgeons and anesthetists are comfortable with performing surgical cricothyroidotomy (Mendonca et al., 2017). This is problematic because confidence enhances the surgical procedure learning process (Peyre et al., 2006). Further, prior research has described how difficult it is to develop procedural competence due to the clinical rarity of cricothyroidotomy (Ferguson et al., 2016). Unsurprisingly, gaining experience with the procedure is associated with higher confidence; however, this relationship appears to plateau such that, after a point, additional experience contributes minimally to increases in confidence (Sergeev et al., 2012). Thus, this procedure requires high amounts of confidence in the capability to complete the procedure successfully, or one's cricothyroidotomy self-efficacy.

Bandura (1977) defined self-efficacy as one's belief in the capability to complete a task successfully. Self-efficacy is critical to self-regulatory systems (Bandura, 1991) involved in skill acquisition. In self-regulatory models, self-efficacy acts as the point of comparison when a task performer judges how much effort they must exert to accomplish set goals. However, there is a downside of high self-efficacy: if individuals believe themselves highly capable, they do not perceive as great a need to exert effort during training (Vancouver, 2008).

Self-efficacy ratings are subject to overconfidence biases that might lead to inaccurate assumptions of confidence. Indeed, it is appropriate to differentiate learner ability to perform surgical cricothyroidotomy (competence) from the learner's perception of cricothyroidotomy self-efficacy (confidence). The relationship between competence and confidence is moderate at best and is contingent on many contextual factors (Zell & Krizan, 2014). The relationship is stronger when the task is objective, familiar, and low in complexity. However, the environments in which one might perform cricothyroidotomy are highly variable and subject to unexpected conditions that may induce unfamiliarity or complexity; such scenarios will likely have a negative impact on one's confidence calibration. Additional researchers found surgical confidence to be related to surgical competency only after training (Clanton et al., 2014), further emphasizing the need for frequent and adequate cricothyroidotomy training. Additionally, multiple studies have observed that self-reported surgical competency from a novice may be inaccurate, unreliable, or both (Clanton et al., 2014; Haddad et al., 2021). Thus, the self-efficacy of novices may be more likely to display evidence of miscalibrated confidence; trainers and researchers should be aware of this and remain even more cautious about inferring competence from novices' self-efficacy ratings. This dangerous inference exists in cricothyroidotomy research that examines changes in confidence as the ultimate outcome as opposed to objective metrics of competence (Zhang et al., 2021).

There are several meta-cognitive biases that affect self-efficacy, such as the Dunning-Kruger effect (Kruger & Dunning, 1999). The Dunning-Kruger effect states that those who are unskilled at a task lack the metacognitive ability to be aware of their own incompetence. So, incompetence reflects not only lower technical skill but less accurate cognitive self-assessment as well. The result is that low performers tend to overestimate their own technical skills (Mahmood, 2016).

Researchers have discussed the implications of the Dunning-Kruger effect in both didactic and medical environments (Gibbs et al., 2017; Rahmani, 2020; Talsma et al., 2019; Winner & Millwater, 2019). Dunning-Kruger effects in didactic training environments may seem benign at first, however the knowledge and confidence gained in such environments is what must transfer to applied medical emergency contexts. In such contexts, human error can threaten patient safety; confidently making an incision in the wrong location can have life-threatening consequences.

Detecting cognitive biases in cricothyroidotomy trainees is critical, as low effort may result from overconfidence. Models of self-regulation state attention is expended *if* learners perceive a discrepancy between current and ideal states, but not if they perceive themselves as already competent enough (e.g., Bandura, 1991; Carver & Scheier, 1982). So, overconfidence likely results in insufficient effort being exerted to perform a task well (Vancouver et al., 2008). Unsurprisingly, any resulting lack of effort has detrimental effects on the rate at which trainees acquire skills.

Addressing the issue of miscalibrated self-efficacy is difficult because many analyses of the Dunning-Kruger effect are influenced by statistical artifacts. As described by Gignac and Zajenkowski (2020), prior methods of detecting the Dunning-Kruger effect categorized continuous variables in a way that fostered information loss. For example, arbitrarily categorizing performance in success quartiles removes any differences in values within such quartiles (i.e., analyses would no longer be able to account for differences between performers in the 76^{th} percentile and those in the 99^{th} percentile). Also, previous findings may have been influenced by the better-than-average effect and regression toward the mean. That is, studies finding evidence of overconfidence may be a result of (a) the tendency to inflate perceptions relative to others (i.e., better-than-average effect) and (b) the statistical tendency of self-perceptions to be biased away from individual ability and toward the mean across multiple individuals (regression toward the mean; Krueger & Mueller, 2002). Simulating statistical artifacts produced results that prior research may have considered supportive of the Dunning-Kruger effect (Gignac & Zajenkowski, 2020). Thus, alternative metrics of the Dunning-Kruger effect should be considered.

Gignac and Zajenkowski (2020) proposed analyses to examine calibration non-linearity and calibration variability. First, they suggested the Glejser test of heteroscedasticity that calculates how variability in an outcome variable (i.e., self-efficacy) may change at differing levels of a predictor variable (i.e., performance; Glejser, 1969). Glejser (1969) described how analysts may regress the absolute value of the residuals from a simple linear regression onto the same predictor variable that produced the residuals to detect heteroscedasticity. Such an analysis reflects the extent to which individuals with low performance produce higher variability in self-efficacy than those with high performance. For example, surgeons who consistently perform at a high level likely all have high self-efficacy, whereas those who either struggle to achieve known benchmarks of performance *or* lack the knowledge of such benchmarks vary in their self-efficacy more. Second, Gignac and Zajenkowski (2020) suggested the use of nonlinear regression to detect a significant change in the relationship between performance and self-efficacy at varying levels of performance.

The Dunning-Kruger effect is thought to generalize to a variety of task types including medical training. When training medics to complete intricate tasks that demand a combination of advanced task knowledge and procedural skill, it is possible for trainees to report mis-calibrated self-efficacy ratings. As with other skilled task performers, medical trainees can hold biases that inflate self-efficacy (e.g., impression management, ego defense). Consequently, cricothyroidotomy trainees are at risk to exhibit evidence of the Dunning-Kruger effect.

An important consideration in cricothyroidotomy training is the effect simulator choice has on confidence development. Prior research has examined this question, finding conflicting evidence such that some studies observe higher confidence in trainees using porcine models compared to a TraumaMan® manikin (Hall, 2011), whereas others observe no difference (Hall et al., 2014). Confidence level aside, trainees still reported preferring to learn using the porcine tracheas (Hall et al., 2014). There are countless forms of cricothyroidotomy simulation with notably distinct physical cues and characteristics, as described earlier. For example, surveys of surgeons and anesthetists found that palpable cricothyroid membrane is associated with higher confidence from anesthetists than surgeons; however, when the cricothyroid membrane was not palpable, such differences disappeared and all respondents were not confident (Mendonca et al., 2017). Such comparisons of confidence development between competing simulators should be maintained and applied to whatever set of simulators one is considering. Further, it is important to maintain awareness of the distinction all simulators share apart from live patient cricothyroidotomy. That is, even if bolstering trainees' confidence, higher fidelity simulators will likely lead to confidence ratings more appropriately aligned with a live patient scenario.

Different forms of cricothyroidotomy may have different patterns of effects on confidence development and retention. Indeed, researchers have examined differences in confidence attainment separated by type of cricothyroidotomy (e.g., cannula, Melker and scalpel-bougie, scalpel-bougie; jet oxygenation; Zhang et al., 2021), finding all techniques led to significant improvements in confidence from pre- to post-training. Further, researchers discovered different longevities of confidence among cricothyroidotomy training types. That is, nine months after training, trainee confidence was maintained for cannula, Melker and scalpel-bougie cricothyroidotomies but not for jet oxygenation methods.

The issue of confidence and competence should be an issue at the forefront of practitioners' minds especially given the high confidence demands and the low frequency of the procedure. That is, there is a "sweet spot," such that trainees should have enough confidence to engage the procedure and perform it well, but not so much that they become overconfident in their abilities. Trainers should frequently consider if those with high cricothyroidotomy self-efficacy are competent or not. Placing an overconfident trainee in a position to perform cricothyroidotomy could have life-threatening consequences. This concern echoes the larger theme of the current paper: evaluating the fidelity of cricothyroidotomy simulators is critical, as training on higher fidelity simulators likely leads to more accurate perceptions of ability relative to real-world cricothyroidotomy.

BEST PRACTICES AND RECOMMENDATIONS

A seminal purpose of the current document is to provide an example of optimal literature review methodologies to complement existing questionnaire-based, functionality evaluation strategies such as those described by Riusi (2019) and Rooney (2019). Reviewing literature represents an opportunity to add to targeted task analytic and data collection approaches for determining simulator realism.

Conducting a literature review is neither complex nor novel. However, the process may seem daunting because of the magnitude and diversity of available information sources that exist across civilian, academic, and governmental archives. Increasingly, such sources include everything from recommendations embedded within formal, peer-reviewed journal articles to user opinions that exist within non-referenced blogs online. The following points are intended to guide procurement efforts. Using surgical cricothyroidotomy as a use case, we first offer recommendations that are general, followed by specific recommendations that originate from our study of surgical cricothyroidotomy simulation devices.

Sources. When considering the sources to consult for a comprehensive literature review, it is important to include governmental, academic, and industrial sources. This process may be rendered more difficult because of the proprietary nature of some governmental and most industrial publications. Ultimately, this may necessitate an investment to ensure that important publications are not overlooked. Fortunately, there are resources online that may alleviate the burden of acquiring reduced-access publications. Those resources include researchgate.com, linkedin.com, and Google Scholar. As for governmental publications, the most widely acknowledged repository for published reports is the Defense Technical Information Center (DTIC).

Time Window. It is clearly difficult to narrow down a literature search by chronological period. When considering constructs of interest (e.g., workload, situation awareness, training efficacy, etc.), some knowledge of the progression of scientific inquiry is helpful. For example, training research began in earnest during the early part of the 20th Century as psychology embraced behavioral methods and simulation grew out of the early Link aviation trainers. Similar events shaped other construct investigations. Toward the end of the 20th Century, the rise in popularity of the internet made acquisition of sources easier across the globe. Consequently, some areas of theoretical important received greater attention from the 1990s to the present.

Weighting of Information. Emphasis given to foundational elements of certain scientific inquiries is best determined by considering the basic and applied importance of the inquiry. As an example, advances in virtual and augmented reality rely directly on knowledge of perceptual activities that include visual, auditory, and tactile processing. However, technical advances associated with computer processing and peripheral design occurred concomitantly with knowledge of perception. To dis-include either is to risk an incomplete review of seminal sources. Other areas (linguistic processing, trust, training, leadership, and others) demand similar comprehensive consideration of sources. Ultimately, the weighting scheme applied to a review must be adjustable to account for advances and gaps in knowledge.

Aspects of Review (Sections). Organization of a literature review may be particularly challenging, especially when the targeted constructs are interrelated. It is helpful to approach such a task by accessing other similar reviews. If none are available, it is helpful to make a master list of constructs, then to subsequently order them in terms of their perceived importance, relation to the overall task goal, or contribution to the literature or the applied goals of the review. Importantly, the structuring of review sections must be malleable and flexible so that they can be rearranged if desired. In some cases, the sections that make up the review may need to be hierarchically ordered to allow for an appreciation of the structural relationships among concepts.

Peer Review vs. Public Access Information. Over the last three decades, the scientific review and publication processes have advanced considerably. Before the 1990s, publication of articles and reports took place in physical format only. Papers were manually produced and replicated, and physically distributed to reviewing entities. Since the 1990s, however, the internet has enabled much quicker review periods and has enabled freer access to research products. At the same time, the quality of published products has varied greatly. In modern times, published reports may be afforded expert review at a high, medium, or low level of rigor. Often, the dividing line between good and poor science is blurred, leaving the responsibility of scrutiny to the ultimate consumer. In some cases, journals or lay outlets basically guarantee publication of a paper for a monetary price. Ultimately, a choice must be made about the types of published works to include and dis-include from inclusion. Perhaps the clearest guidance is to include only those sources that reflect peer scientific review. However, this may conflict with the recency of publication (lightly reviewed works generally achieve publication more quickly).

In addition to the general points made above concerning the steps undertaken during literature review, the following guidelines refer specifically to the consideration of simulation fidelity.

- The highest fidelity simulator may not necessarily be the most effective simulator for training. Many lower-fidelity simulator options have been favorably rated for skill acquisition. Depending on training goals, using a lower fidelity simulator may allow learners to quickly acquire task fundamentals with minimal expenditure of time or funds.
- Though a high-fidelity simulator often facilitates performance improvements, lower-fidelity simulators have been shown to achieve the same goal. In fact, in nearly all the simulator training studies, all models used for training let to some improvement of skill.
- Though study authors or simulator developers may consider their models to be "high-fidelity," such characterizations do not necessarily equate to faithful representation of the target task. The construct of fidelity has been defined in a plethora of ways, leading to confusion in the scientific and lay training communities.
- When considering simulators, it is important to hold important psychological fidelity. Specifically, it is critical to determine whether the simulator prompts appropriate responses throughout the simulation. This relates to the presentation of proper cues *in situ*.
- Ensuring that there are opportunities for feedback is critical for simulation-based training. Failure to do so will jeopardize training effectiveness and learner comprehension.
- Opportunities for repetitive practice are necessary for muscle memory and precise skill development, especially for novice learners.

REFERENCES

Ahn, B., Choi, W. & Jung, J. (2019). Incision Sensor Using Conductive Tape for Cricothyrotomy Training Simulation With Quantitative Feedback. *IEEE Access, 7*, 12947-12958. http://dx.doi.org/10.1109/ACCESS.2019.2891958

Asselin, A., Lafleur, A., Labrecque, P., Pellerin, H., Tremblay, M. H. & Chiniara, G. Simulation of Adult Surgical Cricothyrotomy for Anesthesiology and Emergency Medicine Residents: Adapted for COVID-19. *MedEdPORTAL, 17.* http://dx.doi.org/10.15766/mep_2374-8265.11134

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, *84*(2), 191.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational behavior and human decision processes*, *50*(2), 248-287.

Berwick, R. J., Gauntlett, W., Silverio, S. A., Wallace, H., Mercer, S., Brown, J. M., ... & Groom, P. (2019). A mixed-methods pilot study to evaluate a collaborative anaesthetic and surgical training package for emergency surgical cricothyroidotomy. *Anaesthesia and Intensive Care*, *47*(4), 357-367.

Boet, S., Borges, B. C. R., Naik, V. N., Siu, L. W., Riem, N., Chandra, D., Bould, M. D., & Joo, H. S. (2011). Complex Procedural Skills are Retained for a Minimum of 1 Yr After a Single High-Fidelity Simulation Training Session. *British Journal of Anesthesia, 107*(4), 533-539. http://dx.doi.org/doi:10.1093/bja/aer160

Bowyer, M. W., Streete, K. A., Muniz, G. M. & Liu, A. V. (2008). Immersive Virtual Environments for Medical Training. *Seminars in Colon & Rectal Surgery, 19,* 90-97. https://dx.doi.org/10.1053/j.scrs.2008.02.005

Bryant, R. J., Morgan, M. H., Younquist, S. T. & Fix, M. L. (2017). Edible Cricothyrotomy Model: A Low-Cost Alternative to Pig Tracheas and Plastic Models for Teaching Cricothyrotomy. *Journal of Education and Teaching in Emergency Medicine, 2*(1), 1-6. http://dx.doi.org/10.21980/J8C88S.

Calvo, A., Esteve, C. I., Varela, V., Gomez-Lopez, L., Perdomo, J. M., Berge, R. & Sancho, C. G. (2021). Design, Application and Evaluation of a Cricothyrotomy Model for a Multidisciplinary Simulation. An Observational Single Centre Study. *EducaciónMédica, 22*(4)*,* 305-310. http://dx.doi.org/10.1016/j.edumed.2020.12.003

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological bulletin*, *92*(1), 111.

Carvey, M. M. T., Baek, W. K., Gluschitz, S. & Hage, R. (2020). The Necessity of Practical Emergency Cricothyroidotomy Training During Undergraduate Medical Education. *Translational Research in Anatomy, 19.* https://doi.org/10.1016/j.tria.2020.100070

Cannon-Bowers, J., Bowers, C., Stout, R., Ricci, K., & Hildabrand, A. (2013). Using Cognitive Task Analysis to Develop Simulation-Based Training for Medical Tasks. *Military Medicine, 178*(1)*,* 15-21. http://dx.doi.org/10.7205/MILMED-D-13-00211

Cho, J., Kang, G. H., Oh, Y. M., Choi, H. J., Im, T. H., Yang, J. H., Cho, Y. S., & Chung, H. S (2008). Comparison of Manikin Versus Porcine Models in Cricothyrotomy Procedure Training. *Emergency Medicine Journal, 25*(11), 732-734. https://dx.doi.org/10.1136/emj.2008.059014

Clanton, J., Gardner, A., Cheung, M., Mellert, L., Evancho-Chapman, M., & George, R. L. (2014). The relationship between confidence and competence in the development of surgical skills. *Journal of surgical education*, *71*(3), 405-412.

Cook, T. M., Green, C., McGrath, J. & Srivastava, R. (2007). Evaluation of Four Airway Training Manikins as Patient Simulators Simulators for the Insertion of Single Use Laryngeal Mask Airways. *Anaesthesia, 62*, 713-718. http://dx.doi.org/10.1111/j.1365-2044.2007.05068.x

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J. & Hatala, R. (2013). Comparative Effectiveness of Instructional Design Features in Simulation-Based Education: Systematic Review and Meta-Analysis. *Medical Teacher, 35,* e867-e898. http://dx.doi.org/10.3109/0142159X.2012.714886

Duggan, L. V., Lockhart, S. L., Romano, K. R., Weingart, S. D., Levitan, R. M. & Brindley, P. G. (2017). Front-of-neck airway meets front-of-neck simulation: improving cricothyroidotomy skills using a novel open-access three-dimensional model and the Airway App. *Can J Anesth, 64,* 1079-1081. http://dx.doi.org/10.1007/s12630-017-0926-9

Escher, C., Rystedt, H., Creutzfeldt, J., Meurling, L., Nystrom, S., Dahlberg j., Edelbring, S., Amoroe, T. N., Hult, H., Fellander-Tsai, L. & Abrandt-Dahlgren, M. (2017). Method Matters: Impact of in-Scenario Instruction on Simulation-Based Teamwork Training. *Advances in Simulation, 2*(25). http://dx.doi.org/10.1186/s41077-017-0059-9

Ferguson, I. M., Shareef, M. Z., Burns, B., & Reid, C. (2016). A human cadaveric workshop: one solution to competence in the face of rarity. *Emergency Medicine Australasia*, *28*(6), 752-754.

Friedman, Z., You-Ten, K. E., Bould, M. D. & Naik, V. (2008). Teaching Lifesaving Procedures: The Impact of Model Fidelity on Acquisition and Transfer of Cricothyrotomy Skills to Performance on Cadavers. *Anesthesia & Analgesia, 107*(5), 1663-1669. http://dx.doi.org/10.1213/ane.0b013e3181841efe

Gala, G. G. & Crandall, M. L. (2019). Global Collaboration to Modernize Advanced Trauma Life Support Training. *Journal of Surgical Education, 76*(2), 487-496. https://doi.org/10.1016/j.jsurg.2018.08.011

Gauger, V. T., Rooney, D., Kovatch, K. J., Richey, L., Powell, A., Berhe, H., & Zopf, D. A. (2019). A Multidisciplinary International Collaborative Implementing Low Cost, High Fidelity 3D Printed Airway Models To Enhance Ethiopian Anesthesia Resident Emergency Cricothyroidotomy Skills. *International Journal of Pediatric Otorhinolaryngology, 114,* 124-128. http://dx.doi.org/10.1016/j.ijporl.2018.08.040

Glass, C. C., Parsons, C. S., Raykar, N. P., Watkins, A. A., Jinadasa, S. P., Fleishman, A. & Gupta, A. (2019). An Effective Multi-Modality Model for Single-Session Cricothyroidotomy Training for Trainees. *The American Journal of Surgery, 218,* 613-618. http://dx.doi.org/ https://doi.org/10.1016/j.amjsurg.2018.12.059

Ghosh, A. & Chaudhury, S. (2021). Emergency Cricothyroidotomy Procedure: Lecture on Surgical Anatomy and Briefing of the Procedure Followed by Skill Assessment in First Semester Medical Students. *Rev Arg de Anat Clin, 13*(3), 124-130. Accessible at www.anatclinar.com.ar

Gibbs, S., Moore, K., Steel, G., & McKinnon, A. (2017). The Dunning-Kruger Effect in a workplace computing setting. *Computers in Human Behavior*, *72*, 589-595.

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*, 101449.

Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, *64*(325), 316-323.

Haddad, N., Tsur, A. M., Nadler, R., Glassberg, E., Benov, A., & Chen, J. (2021). Inexperienced but confident: a survey of advanced life support providers and life-saving interventions in the Israel Defense Forces. *Military Medicine*, *186*(Supplement_1), 261-265.

Hall, A. B. (2011). Randomized Objective Comparison of Live Tissue Training versus Simulators for Emergency Procedures. *The American Surgeon, 77*(5)*,* 561-565. http://dx.doi.org/10.1177/000313481107700514

Hall, A. B., Riojas, R. & Sharon, D. (2014). Comparison of Self-Efficacy and Its Improvement After Artificial Simulator or Live Animal Model Emergency Procedure Training. *Military Medicine, 179*(3), 320-323. http://dx.doi.org/10.7205/MILMED-D-12-00446

Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B., & Cook, D. A. (2014). Reconsidering Fidelity in Simulation-Based Training. *Academic Medicine, 89*(3), 387-392. http://dx.doi.org/10.1097/ACM.0000000000000130

Han M, Portnova AA, Lester M, Johnson M (2020) A do-it-yourself 3D-printed thoracic spine model for anesthesia resident simulation. PLoS ONE 15(3): e0228665. https://doi.org/10.1371/journal.pone.0228665

Heard, A., Gorcon, H., Douglas, S., Grainger, N., Avis, H., & Vlaskovsky, P., Toner, A. (2020). Front—of—Neck Airway with Impalpable Anatomy During a Simulated Cannot Intubate, Cannot Oxygenate Scenario: Scalped-Finger-Cannula Versus Scalpel—Finger—Bougie in s Sheep Model. *British Journal of Anesthesia, 125*(2), 184-191. https://doi.org/10.1016/j.bja.2020.04.067.

Huang, J., Ocariz, S., Sims, C. R., Licatino, L. K., & Warner, P. A. (2021). A Novel Approach to Emergency Airway Simulation Using a 3D-printed Cricothyrotomy Task Trainer. *The Journal of Education in Perioperative Medicine, 23*(3), E670. http://dx.doi.org/10.46374/volxxiii_issue3_sims

Hughes, K. E., Biffar, D., Ahanonu, E., Cahir, T., Hamilton, A. & Sakles, J. (2018). Evaluation of an Innovative Bleeding Cricothyrotomy Model. *Cureus 10*(9): e3327. Https://dx.doi.org/10.7759/cureus.3327

Issa, N., Liddy, W. E., Samant, S., Conley, D. B., Kern, R. C., Hungess, E. S., ohen, E. R. & Barsuk, J. H. (2021). Effectiveness of a Simulation-Based Mastery Learning to Train Clinicians on a Novel Cricothyrotomy Procedure at an Academic Medical Centre During a Pandemic: a Quasi-Experimental Cohort Study. *BMJ Open, 11*, e054746. http://dx.doi.org/10.1136/bmjopen-2021-054746

Jing T T, Shah S J, Dadario N B, et al. (April 30, 2021) Evaluation of a Novel Three-Dimensional Emergency Cricothyrotomy Task Trainer. *Cureus 13(*4): e14775. http://dx.doi.org/10.7759/cureus.14775

John, B., Suri, I., Hillermann, C. & Mendonca, C. (2007). Comparison of Cricothyroidotomy on Manikin Vs. Simulator: a Randomised Cross-Over Study. *Anaesthesia, 62*, 1029-1032. http://dx.doi.org/10.1111/j.1365-2044.2007.05219.x

Johnson, Gail, "The Relationship Of Fidelity On Simulation Performance" (2017). Theses and Dissertations. 2243. https://commons.und.edu/theses/2243

Katayama, A., Nakazawa, H., Tokumine, J., Lefor, A. K., Watanabe, K., Asao, T., & Yorozu, T. (2019). A High-Fidelity Simulator for Needle Cricothyroidotomy Training is Not Associated with Increased Proficiency Compared with Conventional Simulators. *Medicine, 98*(8), e14665. http://dx.doi.org/10.1097/MD.0000000000014665

Kei, J., Mebust, D. P., & Duggan, L. V. (2018). The Real Cric Trainer: Instructions for Building an Inexpensive, Realistic Cricothyrotomy Simulator with Skin and Tissue, Bleeding, and Flash of Air. *The Journal of Emergency Medicine, 56*(4), 426-430. https://doi.org/10.1016/j.jemermed.2018.12.023

Kim, J., Park, J.H., & Shin, S. (2016). Effectiveness of simulation-based nursing education depending on fidelity: A meta-analysis. *BMC Medical Education, May 23*, 16:152.

Kozlowski, S. W. J. & DeShon, R. P. (2004). A psychological fidelity approach to simulation-based training: Theory, research, and principles. In E. Salas, L. R. Elliott, S. G. Schflett, & M. D. Coovert (Eds.), Scaled Worlds: Development, validation, and applications (pp. 75-99). Burlington, VT: Ashgate Publishing.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of personality and social psychology*, *82*(2), 180.

Liu, A., Bhasin, Y. & Bowyer, M. (2005). A haptic-enabled simulator for cricothyroidotomy. *Studies in Health Technology and Informatics*, *111*, 308–3133

Mabry, R., Frankfurt, A., Kharod, C., & Butler, F. (2015). Emergency Cricothyroidotomy in Tactical Combat Casualty Care. *Journal of Special Operation Medicine, 15*(3), 11-19. http://dx.doi.org/10.55460/KYGI-F0VF

Mahmood, K. (2016). Do people overestimate their information literacy skills? A systematic review of empirical evidence on the Dunning-Kruger effect. *Communications in Information Literacy*, *10*(2), 3.

Melchiors, J. Todsen, T. Nilsson, P., Wennervaldt, K., Charabi, B., Bottger, M., Konge, L. & von Buchwald, C. (2015). Preparing for Emergency: A Valid, Reliable Assessment Tool for Emergency Cricothyroidotomy Skills. *Otolaryngology-Head and Neck Surgery, 152*(2), 260-265. http://dx.doi.org/10.1177/0194599814556722

Melchiors, J., Todsen, T., Nilsson, P., Kohl, A. P., Bottger, M., Charabi, B., Konge, L. & von Buchwald, Christian. (2016). Self-Directed Simulation-Based Training of Emergency Cricothyroidotomy: a Route to Lifesaving Skills. *European Archives of Oto-Rhino-Laryngology: and Head & Neck, 273*, 4623-4628. http://dx.doi.org/10.1007/s00405-016-4169-0

Mendonca, C., Ahmad, I., Sajayan, A., Shanmugam, R., Sharma, M., Tosh, W., ... & Kimani, P. K. (2017). Front of neck access: a survey among anesthetists and surgeons. *Journal of Anaesthesiology, Clinical Pharmacology*, *33*(4), 462.

Mikkelson, J., Reime, M. H. & Harris, A. K. (2008). Nursing Students' Learning of Managing Cross-Infections – Scenario-Based Simulation Training Versus Study Groups. *Nurse Education Today, 28,* 664-671, Http://dx.doi.org/10.1016/j.nedt.2007.11.003

Miller, G. T., Harris, T., Choi, Y. S., DeLellis, S. M., Nelson, K. & Magee, J. H. (2018). Augmented Reality and Telestrated Surgical Support for Point of Injury Combat Casualty Care: A Feasibility Study. *LNAI. 10915,* 395-405. https://dx.doi.org/10.1007/978-3-319-91470-1_32

Muller, K. L., Facciola, C. A., Monti, J. & Cronin, A. (2020). Impact of Succinct Training on Open Cricothyrotomy Performance: A Randomized, Prospective, Observational Study of U.S. Army First Responders. *Military Medicine, 185,* e1779-e1786. http://dx.doi.org/10.1093/milmed/usaa035

Murray, A. & Konia, M. R. (2010). IERC SimMan/HPS Simulation Scenario Builder. https://aamc-meded.global.ssl.fastly.net/production/media/filer_public/86/d3/86d390ef-1afe-47f9-ab97-404334142af9/ierc_simman_diff_airway.docx

Netto, F. A. C. S., Zacharias, P., Cipriani, R. F. F., de Mello Constantino, M., Cardoso, M. & Pereira, R. A. (2015). A Porcine Model for Teaching Surgical Cricothyroidotomy. *Colégio Brasileiro de Cirurgiões, 42*(3), 193-196. http://dx.doi.org/10.1590/0100-69912015003012

Nguyen, L. H. P., Bankm I., Fisher, R., Mascarella, M. & Young, M. (2019). Managing the Airway Catastrophe: Longitudinal Simulation-Based Curriculum to Teach Airway Management. *Journal of Otolaryngology-Head & Neck Surgery, 48*, http://dx.doi.org/10.1186/s40463-019-0332-0

Paige, J. T., Kozmenko, V., Yang, T., Gururaja, R. P., Hilton, C. W., Cohn, I., Chauvin, S. W. (2009). High-Fidelity, Simulation-Based, Interdisciplinary Operating Room Team Training at the Point of Care. *Surgery, 145*(2), 138-146. http://dx.doi.org/10.1016/j.surg.2008.09.010

Paige, J. B. & Morin, K. H. (2013). Simulation Fidelity and Cueing: A systematic Review of the Literature. *Clinical Simulation in Nursing, 9*(11), e481-e489. http://dx.doi.org/10.1016/j.ecns.2013.01.001

Pandian, V., Leeper, W. R., Jones, C., Pugh, K., Yenokyan, G., Bowyer, M., Haut, E. R. (2020). Comparison of Surgical Cricothyroidotomy Training: A Randomized Controlled Trial of a Swine Model Versus an Animated Robotic Manikin Model. *Trauma Surgery & Acute Care Open, 5,* e000432. http://dx.doi.org/10.1136/tsaco-2019-000431

Petrosoniak, A., Lu, M., Gray, S., Hicks, C., Sherbino, J., McGowan, M. & Monteiro, S. (2019). Perfecting Practice: a Protocol for Assessing Simulation-Based Mastery Learning and Deliberate Practice Versus Self-Guided Practice for Bougie-Assisted Cricothyroidotomy Performance. *BMC Medical Education, 19*, 1-7. https://doi.org/10.1186/s12909-019-1537-7

Peyre, S. E., Peyre, C. G., Sullivan, M. E., & Towfigh, S. (2006). A surgical skills elective can improve student confidence prior to internship. *Journal of Surgical Research*, *133*(1), 11-15.

Proctor, M. D. & Campbell-Wynn, L. (2014). Effectiveness, Usability, and Acceptability of Haptic-Enabled Virtual Reality and Mannequin Modality Simulators for Surgical Cricothyroidotomy. *Military Medicine, 178*(3), 260-264. http://dx.doi.org/10.7205/MILMED-D-13-00365

Qi, D., Petrusa, E., Kruger, U., Milef, N., Abu-Nuwar, M. R., Haque, M., Lim, R., Jones, D. B., Tukseven, M., Demirel, D., Halic, T., De, S., Saillant, N. (2020). Surgeons with Five or More Actual Cricothyrotomies Perform Significantly Better on a VR Simulator. *Journal of Surgical Research, 252,* 247-254. http://dx.doi.org/10.1016/j.jss.2020.03.021

Quick, J. A., MacIntyre, A. D., Barnes, S. L. (2014). Emergent Surgical Airway: Comparison of the Three-Step Method and Conventional Cricothyroidotomy Utilizing High-Fidelity Simulation. *The Journal of Emergency Medicine. 46*(2), 304-307. http://dx.doi.org/10.1016/j.jemermed.2013.08.065

Rahmani, M. (2020). Medical Trainees and the Dunning–Kruger Effect: When They Don't Know What They Don't Know. *Journal of Graduate Medical Education*, *12*(5), 532-534.

Rojas-Munoz, E., Lin, C., Sanchez-Tamayo, N., Cabrera, M. E., Andersen, D., Popescu, V., Barragan, J. A., Zarzaur, B., Murphy, P., Anderson, K., Douglas, T., Griffis, C., McKee, J., Kirkpatrick, A. W., Wachs, J. P. (2020). Evaluation of an augmented reality platform for austere surgical telementoring: a randomized controlled crossover study in cricothyroidotomies. *Npj Digital Medicine, 3(*75). https://dx.doi.org/10.1038/s41746-020-0284-9

Rooney, D.M., Hananel, D.M., Covington, B.J., Dionise, P.L., Nykamp, M.T., Pederson, M., Sahloul, J.M., Vasquez, R., Seagull, F.J., Pinsky, H.M., Sweier, D.G., & Cooke, J.M. (2018). An approach to value-based simulator selection: The creation and evaluation of the Simulator Value Index tool. *Surgery, 163*, 927-932.

Roza, Z.C. (2004): *Simulation Fidelity Theory and Practice: A Unified Approach to Defining, Specifying and Measuring the Realism of Simulations.* Doctoral thesis: Delft University of Technology, Faculty of Aerospace Engineering, Delft.

Ruisi, J., Wier, G., & Ziegenfuss, A. (2017). Standardizing and validating medical simulation equipment procurement across the Air Force Medical Service. *Journal of Defense Modeling and Simulation: Applications, Methodology, and Technology, 14*(4), 421–426.

Sankaranarayanan, G., Odlozil, C. A., Hasan, S. S., Shabbir, R., Qi, D., Turkseven, M., De, S., Funk, G. & Weddle, R. J. (2022). Training on a Virtual Reality Cricothyroidotomy Simulator Improves Skills and Transfers to a Simulated Procedure. *Trauma Surgery & Acute Care Open, 7*, e000826. http://dx.doi.org/10.1136/tsaco-2021-000826

Scali, V., & Espinosa, J. (2020). Critical skills mannequin training for cricothyrotomy. Mathews Journal of Emergency Medicine, 5(2). DOI: https://doi.org/10.30654/MJEM.100

Schebesta, K., Hupfl, M., Rossler, B., RIngl, H., Muller, M. P., & Kimberger, O. (2012). Airway Anatomy of High-Fidelity Human Patient Simulators and Airway Trainers. *Anesthesiology, 116*(6), 1204-1209. http://dx.doi.org/10.1097/ALN.0b013e318254cf41

Schmitz, C. C., Chipman, J. G., Yoshida, K., Vogel, R. I., Sainfort, F., Beilman, G., Clinton, J., Cooper, J., Reilsen, T., Sweet,M. (2014). Reliability and Validity of a Test Designed to

Assess Combat Medics' Readiness to Perform Life-Saving Procedures. *Military Medicine, 179,* 42-48. http://dx.doi.org/10.7205/MILMED-D-13-00247

Scott-Herring, M., Morosanu, I., Bates, J., Batoon, B. (2020). Cut to Air. *American Association of Nurse Anesthesiology Journal, 88*(2), 116-120. Accessible at http://www.aana.com/aanajournalonline

Senuren, C. O., Yaylaci, S., Kayayurt, K., Aldinc, H., Gun, C., Simsek, P., Tatli, O., & Turkmen, S. (2020). Developing Cricothyroidotomy Skills Using a Biomaterial-Covered Model. *Wilderness & Environmental Medicine, 31*(3), 291-297. http://dx.doi.org/10.1016/j.wem.2020.05.003

Sergeev, I., Lipsky, A. M., Ganor, O., Lending, G., Abebe-Campino, G., Morose, A., ... & Glassberg, E. (2012). Training modalities and self-confidence building in performance of life-saving procedures. *Military medicine*, *177*(8), 901-906.

Shaw, M. R. & Hughes, K. E. (2020). High Risk, Low Volume: Evaluation of a Reusable Cricothyrotomy Model in a Paramedic Difficult Airway Training Course. *Air Medical Journal, 39*(5), 380-382. http://dx.doi.org/10.1016/j.amj.2020.05.017

Shefrin, A., Khazei, A. & Cheng, A. (2015). Realism of procedural task trainers in a pediatric emergency medicine procedures course. *Canadian Medical Education Journal, 6*(1), e68-e73. Available at http://www.cmej.ca

Shen, Y., Hananel, D., Zhao, Z., Burke, D., Ballas, C., Norfleet, J., Reihsen, T., Sweet, R. (2016). A New Design for Airway Management Training with Mixed Reality and High-Fidelity Modeling. *Medicine Meets Virtual Reality, 22,* 359-362. http://dx.doi.org/ 10.3233/978-1-61499-625-5-359

Shetty, K., Nayyar, V., Stachowski, E., & Byth, K. (2013). Training for cricothyroidotomy. *Anaesthesia and Intensive Care, 41*(5).

Stokes-Parish, J., Duvivier, R., & Jolly, B. (2019). Expert Opinions on the Authenticity of Moulage in Simulation: a Delphi Study. *Advances in Simulation, 4*(1), 1-10. http://dx.doi.org/10.1186/s41077-019-0103-z

Siu, L. W., Boet, S., Borges, B. C. R., Bruppacher, H. R., LeBlanc, V., Naik, V. N., Riem, N., Chandra, D. B., Joo, H. S. (2010). High-Fidelity Simulation Demonstrates the Influence of Anesthesiologists' Age and Years from Residency on Emergency Cricothyroidotomy Skills. *Anesthesia & Analgesia, 111(*4), 955-90. http://dx.doi.org/10.1213/ANE.0b013e3181ee7f4f

Takayesu, J. K., Peak, D. & Stearns, D. (2017). Cadaver-Based Training is Superior to Simulation Training for Cricothyrotomy and Tube Thoracostomy. *Intern Emerg Med, 12*(1), 99-102. https://dx.doi.org/10.1007/s11739-016-1439-1

Talsma, K., Schüz, B., & Norris, K. (2019). Miscalibration of self-efficacy and academic performance: Self-efficacy≠ self-fulfilling prophecy. *Learning and Individual Differences*, *69*, 182-195.

Umek, N., Hodzovic, I., Damjanovska, M., Cvetko, E., Zel, J., Seliskar, A., Stopar Pintaric, T. (2020). Rescue Oxygenation Success by Cannula or Scalpel-Bougie Emergency Front-of-

Neck Access in an Anaesthetized Porcine Model. *PLOS ONE, 15(*5), e0232510. https://dx.doi.org/10.1371/journal.pone.0232510

Vancouver, J. B., More, K. M., & Yoder, R. J. (2008). Self-efficacy and resource allocation: support for a nonmonotonic, discontinuous model. *Journal of Applied Psychology*, *93*(1), 35.

Varaday, S. S., Yentis, S. M. & Clarke, S. (2004). A Homemade Model for Training in Cricothyrotomy. *Anaesthesia, 59*(10), 1012-1015.

Walsh, R., Heiner, J., Kang, C., Hile, D., & Deering, S. (2013). Emergency Physician Evaluation of a Novel Surgical Cricothyroidotomy Tool in Simulated Combat and Clinical Environments. *Military Medicine, 178*(1), 29-33. https://dx.doi.org/10.7205/MILMED-D-12-00266

Winner, J., & Millwater, T. L. (2019, September). Evaluating human patient simulation fidelity and effectiveness for combat-medical training. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* (Vol. 8, No. 1, pp. 176-180). Sage CA: Los Angeles, CA: SAGE Publications.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111-125.

APPENDIX C - Missing Simulator Cue Cognitive Impact Analysis

INTRODUCTION

Surgical cricothyroidotomy is a complex combat trauma care procedure of last resort precipitated by a situation where airway intubation and ventilation are both impossible to perform. The procedure has been studied for decades; cognitive task analyses have been completed by Demirel and colleagues (2016) prior to the introduction of the Cric-Key instrument kit, and by Bliss and colleagues during the first phase of the current AFRL-funded research. As a part of the most recent task analysis, critical sensory cues were documented by expert performers and task trainers. The resulting task analysis is available from AFRL.

During the data collection activities in the second phase of this research, investigators noted observations from experienced and inexperienced participant trainees. Key among these observations were noted sensory cue presentation deficiencies of the two simulators used for training: the Laerdal Advanced Life Support manikin simulator and the SynDaver Adult Cricothyroidotomy Simulator. The purpose of this document is to document the participants' observations, to associate them with likely cognitive implications, and to propose candidate solutions for simulator designers and military trainers to consider. Though our approach resembles prior cue identification efforts that organize identified cues by task analysis, the results below emphasize cognitive process considerations of missing cues. Ideally, this may help inform knowledge refinements to didactic training curricula as well as physical simulators.

ANALYSIS

In the two tables below, we have included the data observations taken from experienced and inexperienced trainee observations during experimentation. Though participants used the Cric-Key tool set to interact with the simulators, missing cues pertaining to the Cric-Key tool set were not included in our analysis.

Within the third column of each table, cognitive processes most likely impacted by cue omission are identified. The overall set of cognitive processes was informed by existing behavioral theories and models published in the published research domain. The processes include:

- Short-term memory (capacity and duration) (Baddeley & Hitch, 1974)
- Long-term memory (procedural, episodic, semantic) (Atkinson & Shiffrin, 1968)
- Working memory functions (phonological loop, central executive, visuospatial sketch pad) (Baddeley & Hitch, 1974)
- Prospective memory (Sellen et al., 1997)
- Situation awareness (levels, 1, 2, and/or 3) (Endsley, 1995)
- Cognitive workload dimensions (Hart & Staveland, 1988)
- Activity planning (Owen, 1997)
- Prioritization of action (Brown et al., 2015)
- Decision making strategy formation (Klein, 1998)
- Schema matching/expectation (Chitnis et al., 2020)
- Problem representation and solution generation (Newell & Simon, 1972)

It is important to emphasize that the processes listed above may overlap (for example, working memory and schema matching). It is also possible that certain processes have been overlooked or overemphasized. The final column indicates potential solutions for simulator designers and military training personnel to consider.

Table 1. Missing Cues, Associated Sensory Channels, Cognitive Impacts, and Proposed Resolution for Laerdal Simulator

Missing Cue (Source)	Sensory Channel	Cognitive Impact	Resolution/Remedy
Hyoid Bone (I) – Identifies anatomical landmarks	Visual, Tactile	Situation Awareness (1); procedural LTM	3-D Printed Supplement
Skin Marking Ability (I) – Identifies incision site	Visual	Workload; situation awareness (1, 3); planning	AR Cue Overlay
Bones near incision area (I) – Identifies incision site	Tactile	Decision making; working memory; planning	Sim Redesign
Neck skin rigidity (E) – Realistic Cutting Pressure	Tactile	Decision making; working memory	Sim Redesign
Tissue Elasticity (E) – Realistic Cutting Pressure	Tactile	Decision making; procedural LTM	Sim Redesign
Blood flow from incision site (E, I) – Occludes and lubricates incision site	Visual, Tactile	Schema matching; procedural LTM; situation awareness (2, 3)	AR Cue Overlay
Mucus within incision site (E, I) – Occludes and lubricates incision site	Visual, Tactile	Schema matching; procedural LTM; situation awareness (2, 3)	AR Cue Overlay
Subcutaneous fat layer in tracheal region (I) – Additional incision effort	Visual, Tactile	Schema matching; procedural LTM; situation awareness (2, 3)	AR Cue Overlay
Rush of air upon proper bougie placement (I) – Confirms correct tracheal entry	Auditory	Schema matching; working memory; situation awareness (2, 3)	AR Timed Cue Introduction
Cartilage rings in trachea (E, I) – Confirms correct tracheal entry	Tactile	Schema matching; situation awareness (1)	Sim Redesign
Neck skin tenting (I) – Indicates improper bougie introduction	Visual	Problem solving; working memory	AR Cue Overlay

Mist within BVM tube (E, I) – Indicates clear lung access	Visual	Decision making; problem solving; working memory	AR Cue Overlay
Chest sounds (E, I) – Indicates clear lung access	Auditory	Decision making; problem solving; problem solving	AR Timed Cue Introduction
Bilateral chest rise (E, I) – Indicates proper lung ventilation with BVM	Visual	Schema matching; working memory; situation awareness (2, 3)	AR Cue Overlay

Table 2. Missing Cues, Associated Sensory Channels, Cognitive Impacts, and Proposed Resolution for SynDaver Simulator

Missing Cue (Source)	Sensory Channel	Cognitive Impact	Resolution/Remedy
Physical forehead (I) – Indicates proper patient orientation	Visual	Divided attention; planning	AR Cue Overlay
Hyoid Bone (I) – Identifies anatomical landmarks	Visual, Tactile	Situation Awareness (1); procedural LTM	3-D Printed Supplement
Skin Marking Ability (I) – Identify incision site	Visual	Workload; situation awareness (1, 3); planning; prospective memory	AR Cue Overlay
Bones near incision area (I) – Identify incision site	Tactile	Decision making; working memory; planning	Sim Redesign
Neck skin rigidity (E) – Realistic Cutting Pressure	Tactile	Decision making; working memory	Sim Redesign
Blood flow from incision site (E, I) – Occludes and lubricates incision site	Visual, Tactile	Schema matching; procedural LTM; situation awareness (2, 3)	AR Cue Overlay
Mucus within incision site (E) – Occludes and lubricates incision site	Visual, Tactile	Schema matching; procedural LTM; situation awareness (2, 3)	AR Cue Overlay

Rush of air upon proper bougie placement (E, I) – Confirms correct tracheal entry	Auditory	Schema matching; working memory; situation awareness (2, 3)	AR Timed Cue Introduction
Neck skin tenting (I) – Indicates improper bougie introduction	Visual	Problem solving; working memory	AR Cue Overlay
Mist within BVM tube (E, I) – Indicates clear lung access	Visual	Decision making; problem solving; working memory	AR Cue Overlay
Chest sounds (E, I) – Indicates clear lung access	Auditory	Decision making; problem solving; problem solving	AR Timed Cue Introduction
Air resistance (E) – Confirms proper tracheal introduction of air	Tactile	Prioritization; problem solving; procedural LTM	Sim Redesign
Appearance of physical chest (E, I) – Confirms proper ventilation	Visual	Divided attention; planning; prospective memory	AR Cue Overlay

DISCUSSION

The use of part-task simulation has been advocated by training experts for many decades. However, there is general agreement that related task elements should be rehearsed contiguously and continuously (Osgood, 1949). Practical and cognitive considerations result from missing procedural cues. If a simulator does not support critical elements of the procedure, training benefits will suffer and positive transfer of training to actual trauma care situations will be jeopardized (Thorndike & Woodworth, 1901). Consequently, trainees may be likely to commit errors of omission in practice. As showed within Tables 1 and 2, the cognitive mechanisms related to omission reflect anticipation of task duties (planning, prospective memory), recall of past procedural information (long-term memory), representation and organization of current task elements (working memory, prioritization, decision making, problem solving and Levels 1 and 2 situation awareness) and prediction of future consequences (prospective memory, Level 3 situation awareness).

The solutions we propose in Tables 1 and 2 exploit technologies already available. For example, most visual cues may be introduced or supplemented by using augmented reality to overlay visual stimuli onto a physical simulator. Sushareba and Millitello (2021) have demonstrated the utility of this approach. Analogously, missing or insufficient tactile cues may be made available or more salient by integrating 3-D printed anatomical elements within existing simulation platforms. This approach has been achieved by Fontana, Coursen, and Hrabec (2020).

For certain missing cues such as skin tenting, it may be necessary to redesign simulators to reflect task related consequences. Though potentially time consuming and expensive, we believe the training benefits will justify any resource outlays.

REFERENCES

Atkinson, R.C.; Shiffrin, R.M. (1968). *Chapter: Human memory: A proposed system and its control processes*. Psychology of Learning and Motivation. Vol. 2. pp. 89–195.

Baddeley, A.D.; Hitch, G.J.L (1974). Working Memory. *Quarterly Journal of Experimental Psychology*, 4, 302–309.

Brown, M., Shaw, D., Sharples, S., Jeune, I.L., & Blakey, J. (2015). A survey-based cross-sectional study of doctors' expectations and experiences of non-technical skills for out of hours work. *British Medical Journal Open, 5*, e006102.

Chitnis, R., Tulsiani, S., Gupta, S., & Gupta, A. (2020). Efficient binmanual manipulation using learned task schemas. *Proceedings of the IEEE International Conference on Robotics and Automation*.

Demirel, D., Butler, K.L., Halic, T., Sankaranarayanan, G., Spindler, D., Cao, C., Petrusa, E., Molina, M., Jones, D.B., De, S., & de Moya, M.A. (2016). A hierarchical task analysis of cricothyroidotomy procedure for a virtual airway skills trainer (VAST) simulator. *American Journal of Surgery, 212*(3), 475-484.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*. 37(1), 32–64.

Fontana, J., Coursen, J., Hrabec, D., & Krupp, S.S. (2020). Three-dimensional printed model for cricothyroidotomy training. *Academic Emergency Medicine*, *27*, S334.

Hart, S.G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, Peter A.; Meshkati, Najmedin (eds.). *Human Mental Workload*. Advances in Psychology. Vol. 52. Amsterdam: North Holland. pp. 139–183.

Klein, G. (1998). *Sources of power: How people make decisions*. MIT Press, Cambridge, Mass.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. New York, NY: Prentice-Hall.

Osgood**,** C. E. **(***1949*). The similarity paradox in human learning: a resolution. *Psychological Review, 56*(3), 132–143.

Owen, A.M. (1997). Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives". *Progress in Neurobiology*, 53(4), 431–50.

Sellen, A. J., Louie, G., Harris, J. E., & Wilkins, A. J. (1997). What brings intentions to mind? An *in situ* study of prospective memory. *Memory*, 4, 483-507.

Sushareba, C.E., & Militello, L. (2021). Demonstration: Virtual Patient Immersive Trainer to train perceptual skills using augmented reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Thorndike, E. L. & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions., *Psychological Review,* 8.

APPENDIX D - Minimal Data Method

INTRODUCTION

Often, medical purchasing agents have limited time and resources to fully investigate training simulator options. For instances in which resources are scarce, there is a need for a quick method of generating data to inform decisions. Prior methods of quickly determining simulator utility have addressed a number of considerations for simulator purchasers such as purchase cost, scalability, and reputation of manufacturer (Rooney et al., 2018). Other methods address the functionality of simulators, listing whether the functionality exists or not (Ruisi, 2017). Such methods have considerable value in highlighting logistical concerns and the existence of various functionalities. However, the tools advocated by Rooney and colleagues (2018) and Ruisi (2017) neglect to consider the veridicality with which training devices represent target procedures. That is, prior methods do not allow decision makers to consider the realism (fidelity) of procedure representations. Further, such prior methods did not involve the use of cognitive task analysis to identify the cues that need simulating. Typically, checklist tools to support procurement rely on subjective judgments of a simulator's support for various procedures. Not only does such a focus neglect the quality of procedure representation, it may also rely on subjective judgments about functionality. Such judgments often reflect psychological biases (e.g., confirmation bias, availability heuristics, etc.).

In the paragraphs below, we present a supplemental approach that integrates fidelity as a driving influence. Fidelity is a valuable differentiator for simulators that we believe can be leveraged to enable better decision-making. Our proposed approach should supplement existing methods of functionality determination (e.g., Rooney et al., 2018; Ruisi, 2017), incorporating physical and functional fidelity with respect to both the existence and salience of physical cues, as derived through cognitive task analysis.

BACKGROUND

Simulation fidelity is the degree of similarity between a simulator and what it is meant to represent in reality (Roza, 2004). It is important to measure fidelity because it enables training efficacy; it is counterproductive to purchase a simulator that fails to simulate.

Physical fidelity is the extent to which the simulator duplicates the necessary cues to enable successful performance (Gross & Freeman, 1997; Gross et al., 1998; Hays & Singer, 1989; Liu et al., 2008; Meyer, 1998; Paige & Morin, 2013). Such cues either exist or not in the simulation, but even the ones that exist are sometimes difficult to perceive or interact with. This highlights the distinction between simulator cue existence and salience.

Simulator purchasers should always value physical fidelity because without assessing it, one risks purchasing a simulator that insufficiently replicates physical cues necessary to enable performance on the simulated task. As such, purchasers should consider physical fidelity, with respect to both the existence of cues and the salience.

METHOD

Step 1: Define fidelity requirements (per learning objectives and/or cognitive task analysis)

Determining what cues need to be presented by a simulator can be difficult because of varying mission parameters, trainee skill levels and necessary task elements. Therefore, fidelity requirements should reflect learning objectives identified by an analysis of the target task. This process involves identifying each cue presented by the simulator that enables performance. Task analysis is an informed process whereby iterative listing is made of task steps and associated knowledge, skills and abilities. Though time consuming, task analysis helps purchasers understand important aspects of skill learning. Several task analyses are publicly available and would save considerable time (i.e., through journal publications, technical reports, etc.). If existing task analyses do not exist or are inaccessible, see Crandall and colleagues (2006) or Klein (2000) for a description of task analysis and/or how to conduct one. This includes both elements trainees interact with and elements they do not. Both types of cues can influence performance and should be considered when determining simulator fidelity. See Figure 1 for an example task analysis.

Figure 1

Example Task Analysis Excerpt

Task Element	Time Required	Sensory Cues	Knowledge Required	Possible Errors
G. 1. Position patient with medic's dominant hand on patient head, non-dominant had toward patient's feet.	10 secs	VISUAL: Orientation of patient relative to medic	Relative position of patient and medic	Failure to properly assume position
G. 2. Confirm location, accessibility of needed equipment in medic bag.	5 secs	VISUAL: Equipment locations	Necessary Equipment for Procedure	Failure to Confirm Equipment location
G. 3. Confirm that balloon holds air with syringe.	10 secs	VISUAL: Balloon inflation	Expected Balloon Shape when Filled	Misinterpreting balloon shape
G. 4. Visually ensure balloon is flat (not inflated).	2 secs	VISUAL: Balloon structure	Expected Balloon Shape When Empty	Misinterpreting balloon shape

Note. The shown task analysis is analyzing surgical cricothyroidotomy using Cri-Key equipment.

Step 2: Differentially weigh fidelity requirements (i.e., determine which are most/least important)

Some cues are more important to be present in a simulator than others. Once a set of cues has been determined, one should prioritize the cues by assigning weights to them. This can be done using any scale the user wishes (e.g., 1-5, 0-10, etc.) but will allow the more important cues to drive the decision-making process more than less important cues.

Step 3: Develop a simulator cue rating worksheet

Both cue existence and salience should be considered when purchasing simulators, and we have found a worksheet is a simple yet effective method to determine both. That is, one might list the identified cues that need to be present in the simulator and create a checklist rating existence and salience of cues, also allowing for additional comments when necessary. Notably, not all cues on the worksheet may be directly accessible, but that is to be expected if one is attempting to determine simulator fidelity before purchasing. Some cues may be assessed visually through pictures and videos posted online or in promotional materials. See Figure 2 for an example simulator cue rating worksheet.

(**Note:** If comparing multiple simulators, one only needs to complete Steps 1-3 once and can use the same results to inform subsequent steps)

Figure 2

Example Simulator Cue Rating Worksheet

Simulator Cue Ratings

Rating Key:

0 – Cue is not available from the simulator
1 – Simulator is minimally realistic for detection of this cue
2 – Simulator is somewhat realistic for detection of this cue
3 – Simulator is considerably realistic for detection of this cue
4 – Simulator is extremely realistic for detection of this cue

Task Element	Task Cue	Rating	Comments
Position patient with medic's dominant hand on patient head, non-dominant had toward patient's feet.	VISUAL: Orientation of patient relative to medic		
Confirm location, accessibility of needed equipment in medic bag.	VISUAL: Equipment locations		
Confirm that balloon holds air with syringe.	VISUAL: Balloon inflation		
Visually ensure balloon is flat (not inflated).	VISUAL: Balloon structure		

Note. This is an excerpt from a Simulator Cue Rating Worksheet designed for surgical cricothyroidomy using Cri-Key equipment.

Step 4: Research publicly available sources to ascertain cue existence and salience

To the degree possible, one should search publicly available information to complete the simulator cue rating worksheet. This can be done using search engines, video platforms (e.g., YouTube.com), published journal articles, developers' websites, and other resources. Such resources may be available in other domains as well.

(**Note**: If comparing multiple simulators, repeat Step 4 for each simulator being compared)

Step 5: Compare requirements to the simulator cue existence and salience gleaned from the research

With a completed simulator cue rating worksheet, one should compare the results of their research to the cue priorities laid out in Steps 1 and 2. One might consider summing the realism ratings from each item on the simulator cue rating worksheet and comparing the summed values across all simulators being compared. Higher scores would generally indicate higher physical fidelity.

COMPATABILITY WITH OTHER METHODS

Ruisi (2017) describes SimPERM, a method of checking necessary functionality for a medical task and consulting a database that might tell a user which simulator has most of the necessary functionalities. Similarly, Cost Utility Analysis (Levin & McEwan, 2000 as applied for simulation by Lapkin & Levett-Jones, 2011) identifies a list of cues that are rated as being either present or not. Our method enhances such approaches by including (a) preference toward using a cognitive task analysis, (b) distinction between dichotomous cue presence and continuous cue salience dimensions, and (c) usage of the simulator cue rating worksheet.

This method supplements Rooney and colleagues' (2018) SVI tool but with (a) a more detailed evaluation of simulator quality that we believe should be weighted more heavily than the SVI currently does, (b) preference toward using cognitive task analysis, (c) focus on physical fidelity (in the existence/salience of cues), and (d) usage of the simulator cue rating worksheet. The summed responses to the simulator cue rating worksheet can serve as an input to supplement the SVI with a more detailed and empirically supported rating for "Quality of tutoring/feedback, sim to learner". Further, we would recommend an SVI user increasing the weighting factor of the category compared to the various other considerations; again, a simulator is not useful if it does not simulate.

References

Atkinson, R.C.; Shiffrin, R.M. (1968). *Chapter: Human memory: A proposed system and its control processes*. Psychology of Learning and Motivation. Vol. 2. pp. 89–195.

Baddeley, A.D.; Hitch, G.J.L (1974). Working Memory. *Quarterly Journal of Experimental Psychology*, 4, 302–309.

Brown, M., Shaw, D., Sharples, S., Jeune, I.L., & Blakey, J. (2015). A survey-based cross-sectional study of doctors' expectations and experiences of non-technical skills for out of hours work. *British Medical Journal Open, 5*, e006102.

Chitnis, R., Tulsiani, S., Gupta, S., & Gupta, A. (2020). Efficient binmanual manipulation using learned task schemas. *Proceedings of the IEEE International Conference on Robotics and Automation*.

Demirel, D., Butler, K.L., Halic, T., Sankaranarayanan, G., Spindler, D., Cao, C., Petrusa, E., Molina, M., Jones, D.B., De, S., & de Moya, M.A. (2016). A hierarchical task analysis of cricothyroidotomy procedure for a virtual airway skills trainer (VAST) simulator. *American Journal of Surgery, 212*(3), 475-484.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*. 37(1), 32–64.

Fontana, J., Coursen, J., Hrabec, D., & Krupp, S.S. (2020). Three-dimensional printed model for cricothyroidotomy training. *Academic Emergency Medicine*, *27*, S334.

Hart, S.G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, Peter A.; Meshkati, Najmedin (eds.). *Human Mental Workload*. Advances in Psychology. Vol. 52. Amsterdam: North Holland. pp. 139–183.

Klein, G. (1998). *Sources of power: How people make decisions*. MIT Press, Cambridge, Mass.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. New York, NY: Prentice-Hall.

Osgood**,** C. E. **(***1949*). The similarity paradox in human learning: a resolution. *Psychological Review, 56*(3), 132–143.

Owen, A.M. (1997). Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives". *Progress in Neurobiology*, 53(4), 431–50.

Sellen, A. J., Louie, G., Harris, J. E., & Wilkins, A. J. (1997). What brings intentions to mind? An *in situ* study of prospective memory. *Memory*, 4, 483-507.

Sushareba, C.E., & Militello, L. (2021). Demonstration: Virtual Patient Immersive Trainer to train perceptual skills using augmented reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Thorndike, E. L. & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions., *Psychological Review,* 8.