Maastricht University

Through the eyes of the physician

Citation for published version (APA):

Szulewski, A. (2019). *Through the eyes of the physician: Expertise development in resuscitation medicine*. [Doctoral Thesis, Maastricht University]. Maastricht University. https://doi.org/10.26481/dis.20190620as

Document status and date:
Published: 01/01/2019

DOI:
10.26481/dis.20190620as

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 20 Apr. 2024

Through the eyes of the physician:

Expertise development in resuscitation medicine

Adam Szulewski

The research reported here was carried out at

in the School of Health Professions Education

Cover art: Dr. Max Montalvo

Through the eyes of the physician:

Expertise development in resuscitation medicine

DISSERTATION

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday June 20th 2019, at 10:00 hours

by

Adam Szulewski

Supervisor:	Prof. dr. J.J.G. van Merriënboer
Co-supervisor:	Prof. dr. A. Gegenfurtner, University of Passau, Germany
Assessment Committee:	Prof. dr. A.B.H. de Bruin (chair) Prof. dr. W.N.K.A. van Mook Dr. A.G. van der Niet Prof. dr. F. Paas, Erasmus University Rotterdam Prof. dr. J. Sweller, University of New South Wales, Sydney, Australia

Contents

Chapter 1

General Introduction

The nature of expertise in medicine has been widely studied but remains incompletely understood. This is due, in part, to the inherent complexity and lack of standardization of many real-world medical encounters (K Anders Ericsson, 2004). The care of acutely ill patients (referred to as resuscitation medicine in this dissertation) is a particularly germane example of this type of complex and unstandardized clinical encounter. The high-stakes nature of resuscitation medicine, its characteristic decision-making-under-pressure, and the ubiquity of acute illness across many medical specialties warrants taking a deep dive into expertise development in this field.

Identifying the physician team leader of a resuscitation case as expert or novice is usually straight-forward even for a modestly-informed observer. However, it is remarkably difficult to describe the specific cognitive and behavioural processes that underlie this expertise – even for the physician him or herself (K. A. Ericsson, 2006). Some of the factors that differentiate experts from novices in resuscitation are related to medical knowledge and procedural skill, however this is likely only part of what defines expertise in the resuscitation of patients during medical crises. At least as important are the abilities to lead a complex team of medical professionals, make efficient and effective decisions, maintain situational awareness, communicate effectively and utilize resources efficiently (Hicks, Bandiera, & Denny, 2008). Together, these complex skills are known as crisis resource management behaviours and they evolved from crew-resource management skills that were first described in the aviation industry (Helmreich, Merritt, & Wilhelm, 1999). Though much can be gleaned from aviation, the factors associated with caring for an acutely ill patient deserve their own special consideration if we are to understand the phenomenon of expertise development in this domain. A clearer understanding of these factors will also help improve educational practices in resuscitation medicine for the next generation of physicians.

Once primarily the domain of emergency medicine and anesthesiology, a basic knowledge of resuscitation skills is now becoming expected across specialties (McMurray, Hall, Rich, Merchant, & Chaplin, 2017). With the advent of competency based medical education, the number of postgraduate specialty programs that require their trainees to demonstrate competency in these skills continues to grow. Unfortunately, resuscitation cases are not common in everyday clinical practice, which creates challenges for trainees and their program directors. Because we lack a clear understanding of how expertise develops in this

domain, much of the teaching and assessment in resuscitation is based upon expert opinion. Clearly, an understanding of the progression along the novice-expert continuum in resuscitation medicine is needed to ensure that teaching and assessment practices in this critical arena are grounded in empirical evidence. The goal of this dissertation is to delve into this continuum and to explore the process and assessment of expertise development in this field. This introductory chapter will provide background for the dissertation's main research questions and the associated studies by first discussing expertise, cognitive load theory, expert cognitive processes and will then identify some theoretical gaps.

Expertise

Generally speaking, expertise can be thought of as reproducibly superior domain-specific performance resulting from experience and deliberate practice (K Anders Ericsson, Prietula, & Cokely, 2007). "In pretty much every area, a hallmark of expert performance is the ability to see patterns in a collection of things that would seem random or confusing to people with less developed mental representations. In other words, experts see the forest when everyone else sees only trees" (A. Ericsson & Pool, 2016). Early studies in expertise showed exactly this. When shown randomly arranged chess pieces on a chessboard, grandmasters were no better than novices in remembering their positions (Chase & Simon, 1973). When shown chess pieces from actual chess games, experts were much better than novices in remembering their positions because the configurations were meaningful to them (Adrianus D De Groot, 1965; Adriaan D De Groot, Gobet, & Jongman, 1996). In medicine, expertise represents a practitioner's ability to effectively action a complex network of clinically meaningful information that is housed within long-term memory (Ambrose, Bridges, DiPietro, Lovett, & Norman, 2010). This evolution of cognitive architecture can be conceptualized as being related to how hard an individual needs to think in order to accomplish a task (i.e. his/her cognitive load) and the way in which that individual's thoughts are organized (i.e. his/her cognitive processes). Like what is seen in other domains, as medical expertise develops, a physician's cognitive load appears to decrease while expert-like cognitive processes increase when confronted with a given clinical problem (Van Merriënboer & Sweller, 2010). Figure 1 provides a graphical representation of this concept.

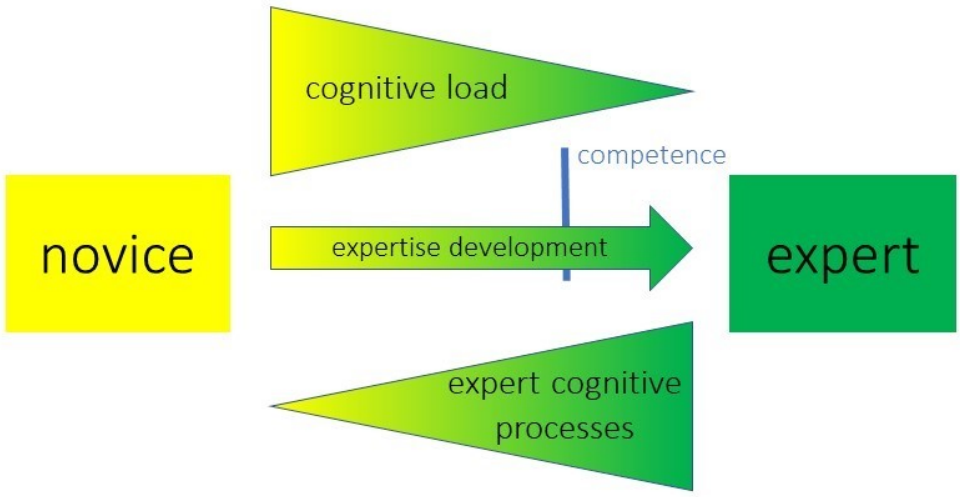


Figure 1: Cognitive load decreases while expert-like cognitive processes increase for a given task as expertise develops in a domain.

Building on this, the present dissertation will examine expertise development in resuscitation medicine through a cognitive load and cognitive process lens. Understanding the progression of expertise in medicine has important implications for understanding the perspectives of medical trainees and education in general (Gunderman, Williamson, Fraley, & Steele, 2001).

Cognitive load theory

Cognitive load theory (CLT) is a theory of education that is grounded in the idea that the human brain is limited by working memory, which is responsible for the ability to process information (Sweller, Van Merrienboer, & Paas, 1998). From a theoretical perspective, cognitive load is thought to be comprised of three basic elements: intrinsic cognitive load, extraneous cognitive load and germane cognitive load. Intrinsic cognitive load refers to the relative complexity of information specific to a task as well as prior knowledge; while

extraneous cognitive load is due to suboptimal information presentation conditions. The sum of intrinsic and extraneous cognitive load is thought to represent the overall cognitive load that can be measured experimentally. Germane cognitive load is thought to refer to the working memory resources dedicated to processing intrinsic cognitive load and is related to the construction and automation of mental schemas (Sweller, 2010). Previous work has demonstrated that these concepts can be measured and change with experience (Leppink, Paas, Van Gog, van Der Vleuten, & Van Merrienboer, 2014). Specifically, cognitive load has been shown to be quantifiable through an analysis of physiologic variables, psychometric surveying as well as secondary task techniques (Laeng, Sirois, & Gredebäck, 2012; Paas, Tuovinen, Tabbers, & Van Gerven, 2003). The reduction in cognitive load (or perceived mental effort) that is apparent in experts is thought to be the result of mental schema construction and automation, which effectively extends working memory by synthesizing interconnected information (Van Merriënboer & Ayres, 2005). In medicine, these schemas are conceptualized as illness scripts that increase in number and become more elaborate as medical expertise develops (Custers, 2015; Schmidt & Rikers, 2007). With the development of expertise and the creation of schemas, physicians alter the way their working memory is utilized as well as their overall cognitive load when dealing with clinical problems. This expert cognitive architecture is likely the basis for some of the variation in expert-specific behaviours observed in physicians of different levels of training.

Of note, though CLT is a theory of learning, many of the principles espoused in CLT seem to be transferable to other related domains like performance, though these parallels have not been fully elucidated in the literature (Plass, Moreno, & Brünken, 2010).

Expert cognitive processes

In addition to changes in cognitive load, researchers have also described measurable changes in expert-like cognitive processes as physicians move along the expertise continuum (Gegenfurtner, Siewiorek, Lehtinen, & Säljö, 2013). These include leadership techniques, crisis resource management skills, as well as eye movement characteristics that change with experience (Blum et al., 2004; Gegenfurtner, Lehtinen, & Säljö, 2011; Kok &

Jarodzka, 2017). In the information reduction hypothesis, Haider and Frensch (1999) describe the tendency of domain experts to disregard task redundant stimuli in their environments, instead focusing on information relevant for the task at hand. This ability to reduce information at a perceptual level leads to greater cognitive efficiencies for the expert and may help to differentiate them from novices.

The way individuals make decisions when dealing with the type of ill-structured problems seen in resuscitation medicine has also been found to change with the development of expertise (Voss & Post, 1988). Decision-making in these contexts is often investigated using a cognitive task analysis approach (Crandall, Klein, Klein, & Hoffman, 2006). The dual-process theory suggests that many of the decisions that require careful consideration by a novice (termed System II processing) become more automated and less effortful for an expert (known as System I processing) (Kahneman, 2011). In crisis situations, experts are thought to be better at recognition primed decision-making – that is, rapidly recognizing complex situations as ones they have seen previously, generating possible courses of action and testing these by mental simulation and finally proceeding with the first reasonable course of action that is not rejected (Klein, 2008). This evolution in the way that experts think allows them to free up cognitive processing resources and allows them to solve problems and reason more efficiently.

Theoretical gaps

Despite these theoretical underpinnings that have been described in other domains, the specific characteristics that change as a medical trainee moves along the novice-expert continuum in resuscitation medicine have not been fully elucidated, suggesting that there is potential for theory building and improvement in medical education in this domain. By focusing on an analysis of cognitive load and expert-like cognitive processes, this dissertation will examine the process of expertise development in resuscitation medicine. The goal is to provide insights into new endpoints that are correlated to the process of expertise development in this field. If found to be reliable markers of expertise, these endpoints may be able to be used in the future to aid trainee assessment decisions and to provide data-driven analysis of expertise development. This would add information to

traditional assessment modalities that are often criticized for being unidimensional and uninformative. Ultimately, this could inform both medical trainees and their supervisors of where a learner lies on the novice-expert continuum in a given domain. In resuscitation medicine training, this may help to inform assessment decisions with regards to professional clinical activities that can be entrusted to an individual (known as entrustable professional activities or EPAs) within a competency-based medical education framework.

Figure 2 shows the over-arching theoretical framework of the proposed thesis. As learners move along the novice-expert continuum, their cognitive load for a particular task decreases while their cognitive processes become more expert-like. In this dissertation, cognitive load will be measured using a physiologic marker (pupillometry) as well as psychometric surveys. Expert cognitive processes will be characterized by studying visual behaviours with eye-tracking technology as well as interviewing using a cognitive task analysis approach. These data will form the basis to view learner assessment on the novice-expert continuum from a novel perspective, adding to traditional assessment modalities that are currently used in medical education. Figure 3 shows how the theoretical framework fits into the context of increasing task complexity. Specifically, when a learner has demonstrated sufficient competence for a specific task, he/she is thought to have achieved the level of an EPA (entrustable professional activity) and can move forward to the next level of task complexity within a competency based medical education model. Practically speaking, as a medical trainee meets the requirement of an EPA, he/she is deemed to be competent to perform a given set of tasks safely in a clinical setting.

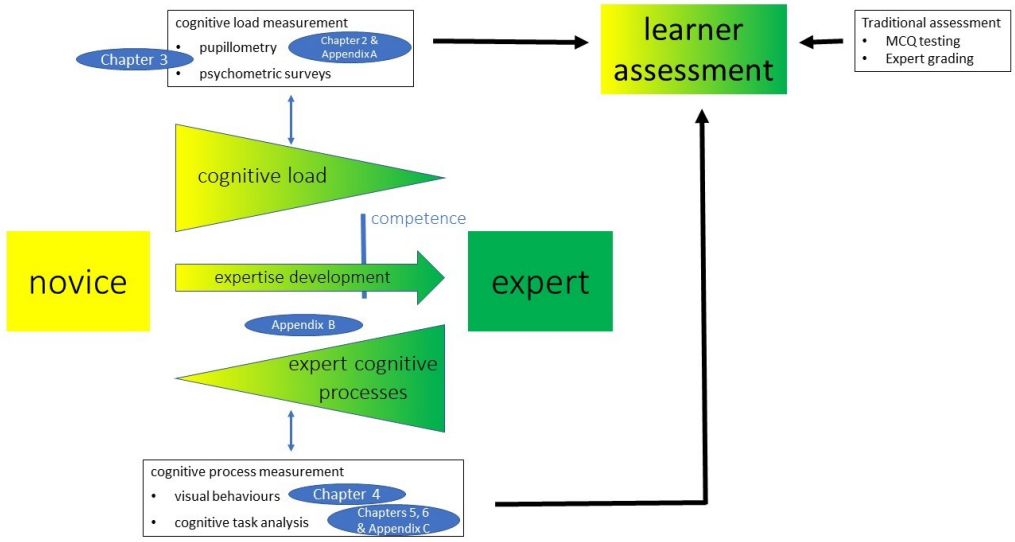


Figure 2: Theoretical framework for the analysis of the development and assessment of expertise in resuscitation medicine. Also shown is where each of the studies conducted fits within this framework.

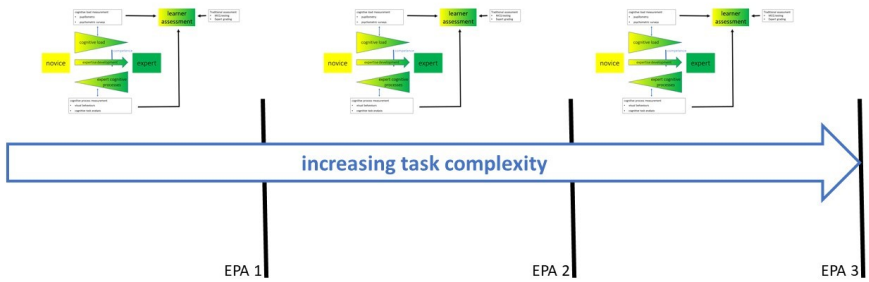


Figure 3: Graphical representation of the theoretical framework within the context of increasing task complexity. EPA refers to entrustable professional activity

The goal of this dissertation is to explore the process and assessment of expertise development in-depth in resuscitation medicine. The common practice of simply observing that a physician is effective during a crisis medical situation does not provide meaningful information about what it means to be an expert, or what educators can do to support the development of expertise. By carefully exploring resuscitation medicine expertise, we will be better able to measure it and assess where learners lie on the expertise development continuum. This is especially relevant in a competency based medical education (CBME) model for resuscitation skills where learners are entrusted with the care of some of the sickest patients in the hospital once they demonstrate a specified level of competency.

With this in mind, the overarching research question of this dissertation is as follows:

> *What changes in cognitive load and processing occur in physicians as they develop expertise within the domain of resuscitation medicine? How can these changes be measured as physicians progress along the expertise continuum?*

Figure 2 provides a graphical representation of how each of the research studies described fits into the overarching theoretical framework. Each numbered study corresponds to the numbered chapters and appendices described below.

Dissertation overview

The following research questions are addressed in this dissertation.

1. What is the relationship of cognitive load, as measured by pupillometry, to level of experience in the context of a traditional knowledge-based resuscitation medicine examination?
2. In the context of a resuscitation medicine examination, what is the validity of using a physiologic measure of cognitive load (pupillometry) and a psychometric one (Paas, 1992) as markers of cognitive load among physicians with different levels of experience?

3. In resuscitation-based simulation OSCE's, how (and to what extent) are particular gaze patterns of residents associated with exam performance? How do these gaze patterns vary across scenarios?
4. What are the cognitive processes of expert physicians while leading actual trauma resuscitations?
5. What are the cognitive processes of medical trainees in simulation-based resuscitation examinations? How do these cognitive processes vary with examination performance?

Chapters 2, 3, 4, 5 and 6 answer each of the research questions sequentially in the form of empirical studies. Appendices A, B and C represent additional work related to this dissertation that is closely related to the contents of the chapters but don't specifically address the research questions themselves. Chapters 2, 3 and Appendix A of this dissertation focus on cognitive load measurement; while Chapters 4, 5, 6 and Appendices B and C deal with cognitive process measurement.

Cognitive load measurement

The first empirical study is presented in Chapter 2, which answers the first research question. In this chapter, pupillometry is used to measure the overall cognitive load of participants of varying levels of experience while completing a traditional examination. Chapter 3 takes this further by providing evidence of construct validity of using both pupillometry as well as a psychometric survey to measure the cognitive load of physicians in a similar testing environment. Doing so, it answers the second research question. Appendix A provides background about the use of pupillometry as a tool to study cognitive load and expertise in medicine in the form of a review paper.

Cognitive process measurement

Chapter 4 answers the third research question by using eye-tracking technology to determine how resident visual behaviours vary with performance on a resuscitation-based

examination in a medical simulator. To understand the cognitive processes of experts in resuscitation medicine in the real-world, Chapter 5 introduces the technique of cognitive task analysis augmented by eye-tracking and answers the fourth research question. An overview of decision-making in crisis situations and its development in individuals over time is provided in Appendix B. The final research question is answered in Chapter 6 where the cognitive task analysis augmented by eye-tracking approach is brought back to the simulation laboratory where the development of expert-like cognitive processes in residents is investigated. Appendix C describes the utility of a new simulation debriefing method that was borne out of the study in Chapter 6.

The dissertation concludes with a discussion chapter that provides a bird's eye overview of the work conducted, theoretical and practical implications of the results, and suggestions for future work in the field.

References

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*: John Wiley & Sons.

Blum, R. H., Raemer, D. B., Carroll, J. S., Sunder, N., Felstein, D. M., & Cooper, J. B. (2004). Crisis resource management training for an anaesthesia faculty: a new approach to continuing education. *Medical Education, 38*(1), 45-55.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55-81.

Crandall, B., Klein, G., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*: Mit Press.

Custers, E. J. (2015). Thirty years of illness scripts: Theoretical origins and practical applications. *Medical Teacher, 37*(5), 457-462.

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

De Groot, A. D., Gobet, F., & Jongman, R. W. (1996). *Perception and memory in chess: Studies in the heuristics of the professional eye*: Van Gorcum & Co.

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise*: Houghton Mifflin Harcourt.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70-S81.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. *The Cambridge handbook of expertise and expert performance*, 223-241.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7/8), 114.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552.

Gegenfurtner, A., Siewiorek, A., Lehtinen, E., & Säljö, R. (2013). Assessing the quality of expertise differences in the comprehension of medical visualizations. *Vocations and Learning, 6*(1), 37-54.

Gunderman, R., Williamson, K., Fraley, R., & Steele, J. (2001). Expertise: implications for radiological education. *Academic Radiology, 8*(12), 1252-1256.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 172.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology, 9*(1), 19-32.

Hicks, C. M., Bandiera, G. W., & Denny, C. J. (2008). Building a Simulation-based Crisis Resource Management Course for Emergency Medicine, Phase 1: Results from an Interdisciplinary Needs Assessment Survey. *Academic Emergency Medicine, 15*(11), 1136-1143.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456-460.

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education, 51*(1), 114-122.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry A Window to the Preconscious? *Perspectives on Psychological Science, 7*(1), 18-27.

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32-42.

McMurray, L., Hall, A. K., Rich, J., Merchant, S., & Chaplin, T. (2017). The Nightmares Course: A Longitudinal, Multidisciplinary, Simulation-Based Curriculum to Train and Assess Resident Competence in Resuscitation. *Journal of Graduate Medical Education, 9*(4), 503-508.
Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429.
Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.
Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*: Cambridge University Press.
Schmidt, H. G., & Rikers, R. M. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education, 41*(12), 1133-1139.
Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138.
Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.
Van Merriënboer, J. J., & Ayres, P. (2005). Research on cognitive load theory and its design implications for e-learning. *Educational Technology Research and Development, 53*(3), 5-13.
Van Merriënboer, J. J., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education, 44*(1), 85-93.
Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In *The nature of expertise.* (pp. 261-285). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Chapter 2

The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise

Published as: Szulewski, A., Roth, N., & Howes, D. (2015). The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: a new tool for the assessment of expertise. *Academic Medicine*, *90*(7), 981-987.

Abstract

Objective
Task-evoked pupillary responses (TEPRs), or changes in pupil size, correlate with changes in cognitive processing demands. The magnitude of this change is a reliable marker of cognitive load. The authors used TEPRs to compare cognitive load between novices and trained physicians as they answered clinical knowledge questions.

Methods
In 2013, twenty emergency medicine trainees were recruited and divided into novice (n = 10) and trained physician (n = 10) groups. The authors used mobile eye-tracking glasses to assess changes in pupil diameter as participants answered arithmetic questions, general knowledge questions, and clinical emergency medicine questions in a controlled setting. Questions were categorized by difficulty a priori.

Results
Difficult arithmetic questions caused greater changes in TEPRs than easy ones ($P = .024$). TEPRs were similar between groups when answering general knowledge questions ($P = .383$) but were significantly greater for novices than trained physicians when answering clinical questions ($P < .001$). TEPRs in trained physicians were significantly greater when answering difficult clinical questions than easy ones ($P < .001$), whereas TEPRs in novices were similar ($P = .291$). For those clinical questions answered correctly by both groups, TEPRs in novices were greater than those in trained physicians despite all participants answering correctly ($P < .001$).

Conclusion
Novices require more mental effort to answer clinical questions than trained physicians, even when both respond correctly. Measuring TEPRs has the potential to be a valuable assessment tool by providing objective measures of expertise and is worthy of further study.

Introduction

Accurate, efficient, and expedient decision-making is essential for providing high-quality patient care in an emergency medicine (EM) setting. Despite ongoing work in this area, decision-making in medicine remains a difficult process to study (Ericsson, 2004).

With experience some clinicians develop *expertise*, defined as measurable, superior performance in a domain, which results from years of experience and deliberate practice (Ericsson, Prietula, & Cokely, 2007). Given its inherent complexity and lack of standardization, expertise in medicine is a challenging phenomenon to quantify, and, as a result, the process of developing medical expertise is not well understood (Ericsson, 2004; Szulewski & Howes, 2014). Comparing the decision-making process of novices to that of experts may offer useful insights and improve our understanding of expertise (Schubert, Denmark, Crandall, Grome, & Pappas, 2013), may provide opportunities to expedite the process of expertise development, and could lead to the creation of more objective measures of competence.

Cognitive load theory posits that the human brain is limited by the capacity of working memory--the component of memory responsible for processing information (De Jong, 2010; Sweller, Van Merrienboer, & Paas, 1998). With increased expertise, clinicians develop more complex and inclusive decision-making strategies (Schubert et al., 2013) and are able to extend their domain-specific, long-term working memory (Ericsson & Kintsch, 1995). As a result, experts use working memory differently than novices, and individual components of clinical work likely cognitively load experts and novices in different ways.

Expertise and the differentiation between novice and expert performance based on cognitive load has been studied in a variety of fields with the use of eye-tracking technology. Research shows that patterns in eye movement fixation and saccadic characteristics are different between experts and novices in a given domain (Gegenfurtner, Lehtinen, & Säljö, 2011). In addition, laboratory research has confirmed that small changes in pupil diameter, on the scale of tenths of millimeters, correlate with changes in cognitive processing demands during tasks such as spelling and solving arithmetic problems and anagrams (Beatty, 1982; Hess, 1965; Hess & Polt, 1964; D. Kahneman & Beatty, 1966; Klingner, Kumar,

& Hanrahan, 2008; Klingner, Tversky, & Hanrahan, 2011; Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Szulewski, Fernando, Baylis, & Howes, 2014). Such measurements have traditionally been imprecise and challenging to gather (DeLeeuw & Mayer, 2008). These task-evoked pupillary responses (TEPRs) occur shortly after the onset of a task and subside quickly after processing is terminated (Beatty, 1982; Klingner et al., 2008; Szulewski et al., 2014). At an anatomical level, these changes are thought to be an involuntary response that results from pathways that originate in the locus coeruleus, a major norepinephrine source in the brain (Brisson et al., 2013).

Studying changes in pupil diameter traditionally required complex and limiting laboratory infrastructure with non-mobile cameras and onerous manual data collection and analysis (Hess, 1965). Newly developed portable devices have facilitated this process, as they allow digital recording of pupil changes and a more convenient means for the quantification of TEPRs in dynamic environments (Szulewski et al., 2014). The utilization of this new technology may allow for the identification of differences in how novices and experts think in medicine. A greater understanding of these differences could lead to the development of practical and successful strategies to expedite the progression from novice to expert physician.

The objective of this study was to use TEPRs to quantify and compare cognitive load among EM trainees with different levels of expertise as they answered clinical EM knowledge questions.

Methods

Participants

Twenty individuals were recruited to participate in our study between July and September 2013. They were divided into two experimental groups based on their level of medical training. The novice group consisted of ten subjects--one first-year medical student and nine second-year medical students. All had successfully completed their respective years of study. The group with more expertise (referred to here as the “trained” group) consisted of ten subjects as well--three fourth-year EM residents, four fifth-year EM residents, and three

EM graduates who had successfully passed their specialty examinations the previous year. We made the decision to include senior residents and junior attending physicians instead of more experienced clinicians in the trained group to limit unfamiliarity with multiple-choice testing as a confounding factor.

We did not offer incentives to participate. Each participant was contacted by a peer at a similar level of training, and the study participation rate was 100% of those approached. The Queen's University Faculty of Health Sciences Research Ethics Board approved this study (SMED-115-13; File # 6010680). We obtained written informed consent from each participant prior to the study.

Experimental design

We presented participants with four multiplication problems, four general knowledge multiple-choice questions (MCQs), and twelve clinical EM MCQs (see Appendix 1). Between questions, a screen showing a black circle was displayed for five seconds to re-establish a pupil diameter baseline. Intermittently, we presented participants with a black star instead of a question, and the pupil size measured during these times served as the control. Participants verbally answered all questions while looking straight ahead at an LED television screen situated one meter away at a constant brightness.

The multiplication questions were of varying difficulty, and each displayed for 10 seconds. The purpose of these questions was to ensure that pupillary responses correlated with cognitive load, as observed in previous studies (Klingner et al., 2011; Szulewski et al., 2014).

The general knowledge MCQs also were displayed for 10 seconds each. We prepared these questions, which consisted of socio-geographical trivia that we assumed would be equally familiar to both groups of participants. The purpose of these non-clinical questions was to discriminate between cognitive load differences due to knowledge base and those based on fundamental differences in cognitive processing ability among the participants.

The clinical MCQs were displayed for 30 seconds and were taken from two published EM review texts--one at the medical student level ("easy questions") (Rosh, 2012) and the other at the resident level ("difficult questions") (Promes, 2013). These practice questions are

widely used in the Canadian medical education system in preparation for examinations at the senior medical student and senior EM resident levels respectively, thus we assumed that they represent different levels of question difficulty. All participants encountered the questions in the same order.

Pupil data recording and analyses

We sampled pupil diameter monocularly at 30 Hz using the Tobii Glasses Eye Tracker (Tobii Technologies, Danderyd, Sweden) and expressed that measure as a percentage change from baseline. We measured the raw data in millimeters and changes in pupil size on the magnitude of tenths of millimeters. The device was calibrated in the same windowless room for each participant according to manufacturer recommendations before the experiment, at which time we determined baseline pupil size (Tobii, 2012). We averaged subsequent pupil diameter measurements for each participant to determine a mean pupillary size reading per second. These data points then were averaged for all participants in each group (novice and trained) to give a composite mean percent pupil diameter change from baseline for each second and each question type. We analyzed pupil diameter changes in both groups for difficult compared with easy questions as well as difficult and easy questions compared with the black star control. To assess the difficulty and item discrimination, we completed a test item analysis on the MCQs (Lange, Lehmann, & Mehrens, 1967; Siri & Freddano, 2011). Next, we determined item difficulty by dividing the number of participants answering the question correctly by the total number of participants. By subtracting the novice item difficulty from the expert item difficulty, we determined the item discrimination index. We decided *a priori* that we would exclude a MCQ from analysis if the percentage of participants who answered that question correctly was less than 30%, the item discrimination index was less than 0.30, or the question was a negative discriminator.

We report data as mean ± standard deviation. Comparisons for change in pupil diameter from baseline were made by unpaired t-test or one-way ANOVA with subsequent post-hoc tests based on the number of groups being compared (Graph Pad Software Inc., La Jolla, CA). We considered differences significant at the $P < .05$ level.

Results

Following the test item analysis, one clinical MCQ was excluded from the data set because the item discrimination index was below the pre-determined value capable of differentiating levels of expertise. All remaining questions were included.

Pupillary responses

Figure 1 shows the change in pupil diameter when answering easy and more difficult arithmetic questions, compared with the change when looking at the control screen. The increase in pupil diameter was significant for all twenty participants when answering both easy (mean = 2.31%, SD = 1.53) and difficult (mean = 6.50%, SD = 3.26) arithmetic questions compared with the control (mean = -0.31%, SD = 0.44) (P = .035 and P = .009, respectively), and for difficult compared with easy arithmetic questions (P = .024). *Post-hoc* analysis showed that these findings were consistent for both the trained group and the novice group.

Figure 2 shows the change in pupil diameter for novices (mean = 2.79%, SD = 1.66) and trained physicians (mean = 2.31%, SD = 1.64) when answering general knowledge questions compared to looking at the control screen (mean = -0.31%, SD = 0.44). Both groups showed an increase in pupil diameter when answering questions compared to viewing the control stimulus (novice group P = .019, trained group P = .026); however, as expected, this response was similar for the novice group compared with the trained group (P = .383).

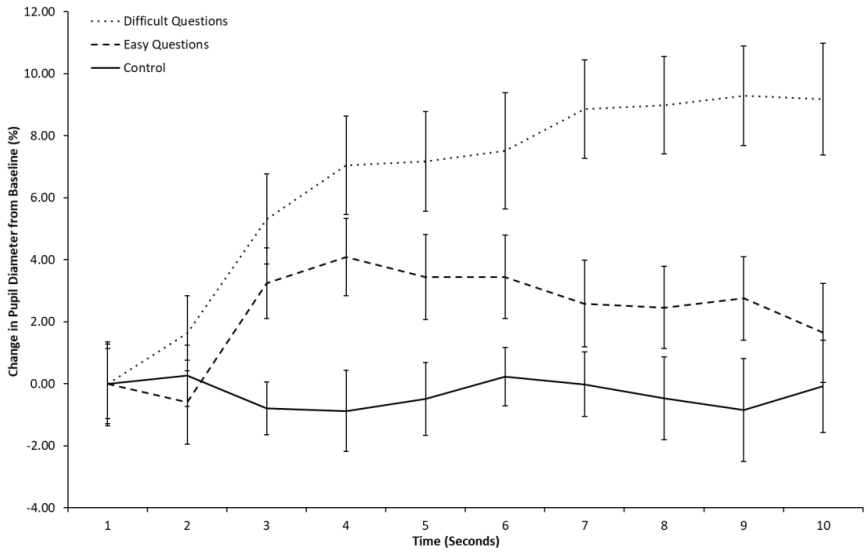


Figure 1: Change in pupil diameter when answering easy and more difficult arithmetic questions, compared to looking at the control screen. Change in pupil diameter was greater when answering difficult arithmetic questions compared with easy arithmetic questions (*P* = .024). The increase was significant for both easy and difficult questions when compared with the control (*P* = .035 and *P* = .009, respectively).

The pooled analysis of the TEPRs for the clinical MCQs (see Figure 3) showed the change in pupil diameter from baseline was significantly greater in novices (mean = 6.28%, SD = 3.43) compared with trained physicians (mean = 2.62%, SD 1.79) over the course of the test (*P* < .001). The TEPRs of trained physicians were significantly different for easy questions compared with difficult questions (*P* < .001), but the TEPRs of novices were not (*P* = .291).

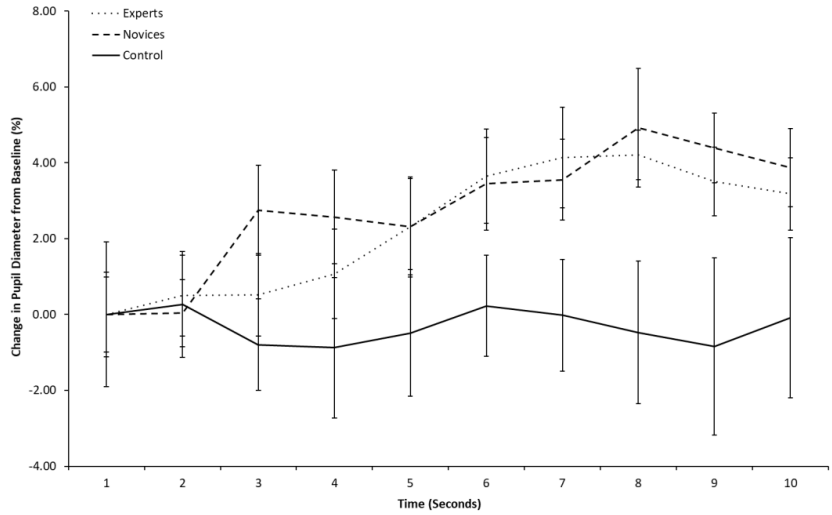


Figure 2: Change in pupil diameter for novices and trained physicians when answering general knowledge multiple choice questions compared to looking at the control screen. No difference was found in pupil diameter between novices and trained physicians when answering these questions (*P* = .383). Both groups showed an increase in pupil diameter when answering questions compared to viewing the control screen (novice group *P* = .019, trained group *P* = .026).

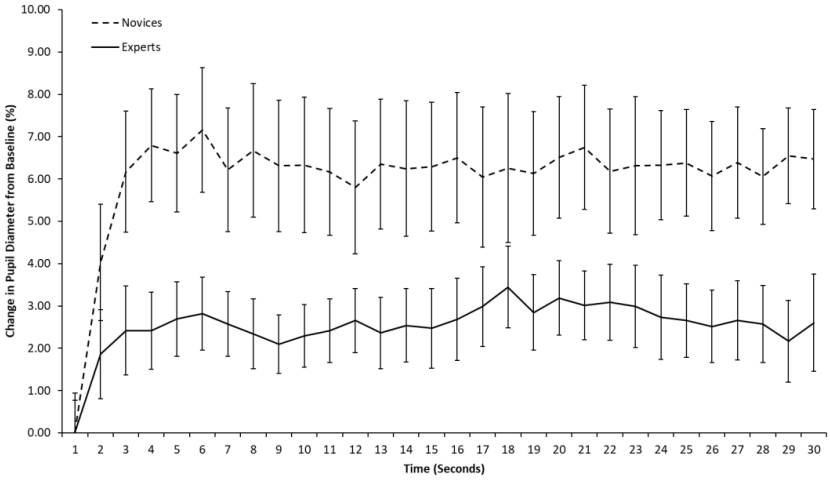


Figure 3: Change in pupil diameter from baseline when answering clinical multiple choice questions. The change was greater in novices compared with trained physicians (*P* < .001).

We analyzed the relationship between TEPRs and the correctness of answers post hoc (see Figure 4). When our analysis was limited to the questions that participants answered correctly, the increase in the pupil diameter of novices (mean = 4.87%, SD = 1.45) was significantly greater than the increase in pupil diameter of trained physicians (mean = 2.54%, SD = 0.56) ($P < .001$).

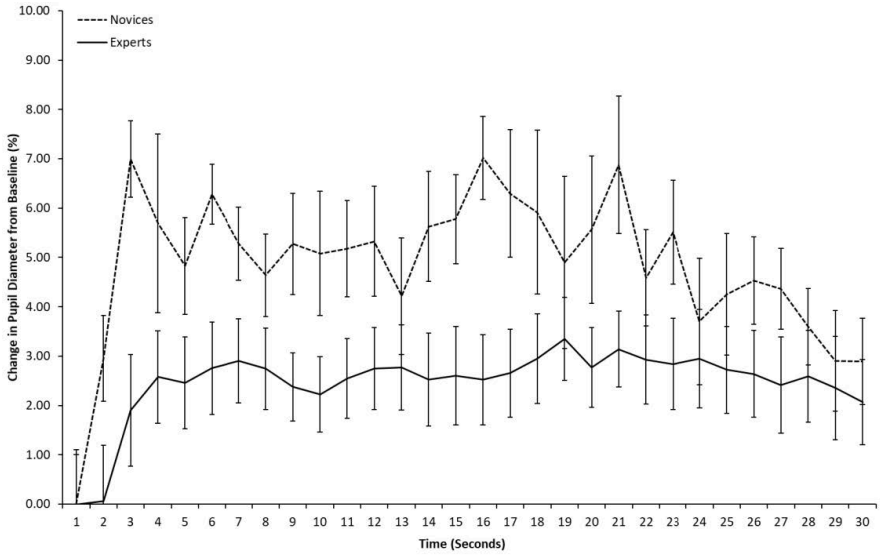


Figure 4: Results of an analysis of correctly answered clinical multiple choice questions. The increase in the pupil diameter of novices was significantly greater than that of trained physicians ($P < .001$).

Effectiveness of cognitive load manipulation

We also analyzed the time it took for participants to verbalize an answer to each MCQ. Trained physicians answered faster than novices for difficult questions (24.1 seconds compared with 28.2 seconds; $P = .001$) as well as for easy questions (16.3 seconds compared with 22.2 seconds; $P = .014$).

Trained physicians answered a mean of 90% of the MCQs correctly, while novices answered a mean of 35% correctly. Overall, participants answered easy questions correctly more often

than they did difficult questions (70% of easy questions answered correctly compared with 48% of difficult questions answered correctly).

Discussion

Our findings demonstrate that pupillometry can be used to objectively measure cognitive load in a testing situation. They also suggest that physicians with more training and experience exhibit less cognitive load than novices when answering questions in their field of expertise.

TEPRs from participants answering arithmetic questions (Figure 1) followed a pattern consistent with that identified in previous studies of cognitive load (Beatty, 1982; Klingner et al., 2008; Szulewski et al., 2014). That is, easy questions evoked a significantly greater response than the control did. The difficult questions produced an even greater response, with the peak TEPR occurring after question presentation, compared with the easy questions. The decrease in cognitive load after the peak response indicates that the subject either answered or gave up on the question (Daniel Kahneman, 2011). Finally, the slight decrease in TEPR from baseline when looking at the control, seen in Figures 1 and 2, may reflect the relaxation that occurred when the participant realized that there was no question requiring a response.

The comparison of TEPRs from trained physicians and novices suggests that physicians answering questions in their field of expertise experience significantly less cognitive load. We found no difference between the groups when they were answering questions outside their field, suggesting that the observed difference is due to an integration of the subject matter rather than to a difference in the baseline cognitive capabilities of the two groups. This findings also suggests that at least some of the cognitive efficiency that comes with expertise in a domain is not transferable to other, unrelated domains. This finding is consistent with the expertise literature that suggests that expert performance is in fact domain specific (Farrington-Darby & Wilson, 2006).

The difference in TEPRs between trained physicians and novices was consistent even when we included only the data for correctly answered questions (Figure 4), further supporting a

fundamental difference in how experts access information that they know well. Thus, one could differentiate a novice from someone with more expertise despite the fact that both may provide the same correct answer to a MCQ. Educators often criticize exams for testing trainees on clinically irrelevant material, a result of the need for the exam to have some discriminatory power. If confirmed with larger studies that examine individuals over time, objective physiologic measures of cognitive function could obviate this practice.

The greater cognitive load experienced by novices when accessing knowledge may help to explain some of the difference in cognitive functioning between novices and experts in the clinical environment. Further exploration of these differences in cognition in simulated and real medical emergencies is needed if we are to understand what learners are experiencing in these situations and how we can best help them to perform better.

We are just beginning to understand the complexity of how expertise develops. In the educational literature, expertise is conceptualized as an increasingly complex network of pieces of information about a topic, growing in number and becoming more interconnected as expertise develops (Ormrod, 2012). Experts establish a sufficiently vast network of concepts in their long-term memory so that their working memory is relatively freed up to think critically and analyze new information (Ambrose, Bridges, DiPietro, Lovett, & Norman, 2010). This evolution of the thinking process allows experts to expend less mental effort solving routine problems, which is consistent with our inherent preference to avoid cognitively demanding executive processing tasks in daily life (Kool, McGuire, Rosen, & Botvinick, 2010).

Our study has a number of limitations. Multiple factors can cause changes in pupil size, including the light and accommodation reflexes. Our experimental design allowed us to control the environment, including lighting and focal distance, and we used a control stimulus to address other potential confounders. The use of corneal reflection eye-trackers, like the one we used in this study, have the potential to lead to errors in the measurement of pupil size, especially when a subject looks away from the camera (Brisson et al., 2013). Despite this potential limitation, one would expect general trends between novices and trained physicians to hold true as both groups were asked the same questions and presumably their gaze drifted away from the center of the camera for similar periods of time as they read each question presented on the screen.

Our analysis used pooled data to identify patterns, and we only examined our participants at a single point in their development of expertise. This study design was useful in identifying trends, but further study will be required to better understand the cognitive load patterns of individuals and how an individual's cognitive processing changes over time. Moreover, we used a convenience sample of individuals as opposed to a random sampling method, which may limit the generalizability of our findings; it also underscores the importance of confirming our findings with additional experiments.

Our study examined cognitive load during a multiple choice exam, isolating one aspect of clinical medicine and excluding issues of information gathering, sensory overload, communication, and emotional distress. Further studies in real-life complex medical environments may better define other contributors to cognitive load and provide more information about how expert clinicians adapt and change decisions based on new information, manage competing interests, and maintain situational awareness. Ultimately, doing so may help to determine whether experts experience less cognitive load, are better able to deal with high cognitive load, or a combination of the two. This information will provide greater insight into the nature of expertise and has the potential to advance medical education and assessment.

The possibility of objective measures of expertise is exciting. Currently, we rely on objective measures of knowledge combined with subjective measures of performance to assess competence. The objective measurement of expertise will allow us to further study the process of moving from novice to expert, identify developmental milestones in this process, and discern specific barriers experienced by certain individuals. These measures also may allow teachers and researchers to further differentiate between learners who have a superficial knowledge and those with a deeper understanding of the content.

Conclusions

Measurements of cognitive load using TEPRs suggest that novices require more mental effort to answer clinical questions than do physicians with more experience, even when both respond with the correct answer. TEPRs may offer an objective measure of expertise

and warrant further investigation as a potential assessment tool. Continued research is needed to better understand the cognitive maturation process that takes place as a clinician transitions from novice to expert.

Acknowledgements

The authors would like to thank the Kingston Resuscitation Institute for providing access to the eye-tracking device as well as to Wilma Hopman for assistance with statistical analysis.

Ethical approval

This study was approved by the Queen's University Faculty of Health Sciences Research Ethics Board (SMED-115-13; File # 6010680).

Appendix 1

Sample Questions in a Study Using Task-Evoked Pupillary Response to Compare Cognitive Load in Novices and Trained Physicians

8 X 11	Easy arithmetic question
12 X 16	Difficult arithmetic question
Switzerland consistently ranks as offering one of the highest qualities of life in the world. What is the capital of this country? A. Zurich B. Geneva C. Basel D. Bern E. Lausanne	General knowledge question
You are treating an obtunded 72-year-old female for severe sepsis due to pneumonia. You have administered 3 L normal saline and broad-spectrum antibiotics. After performing endotracheal intubation, you place a central venous oxygen saturation catheter and arterial line in preparation for early goal directed therapy. The patient's blood pressure is now 88/34, central venous pressure is 14 mm Hg, and venous oxygen saturation is 66%. Lab values are not yet available. What is the MOST appropriate next step? A. Administer 1 L normal saline intravenously. B. Begin infusion of dobutamine 0.5-1.0 g/kg/min IV. C. Begin infusion of norepinephrine 5 mcg/min IV. D. No further actions, the endpoints of early goal directed therapy have been met. E. Transfuse 2 units packed red blood cells.	Difficult clinical multiple choice question
A 49-year-old male with a history of chronic renal insufficiency is referred to the ED for palpitations and fatigue. His potassium level was found to be 6.9 by his primary doctor. Upon arrival, the patient has a heart rate of 56, a widened QRS on ECG, and a BP of 110/76. He is speaking to you, his lungs are clear bilaterally, and his oxygen saturation is 100% on 2 L NC. He has a large bore IV and is placed on the monitor. Which of the following is the most appropriate initial treatment for this patient with severe hyperkalemia? A. Albuterol nebulizer B. Calcium gluconate C. Insulin and glucose D. Sodium bicarbonate E. Sodium polystyrene sulfonate	Easy clinical multiple choice question

References

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*: John Wiley & Sons.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276-292.

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods, 45*(4), 1322-1331. doi:10.3758/s13428-013-0327-0

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*(2), 105-134.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70-S81.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7/8), 114.

Farrington-Darby, T., & Wilson, J. R. (2006). The nature of expertise: A review. *Applied Ergonomics, 37*(1), 17-32.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552.

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

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583-1585.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). *Measuring the task-evoked pupillary response with a remote eye tracker.* Paper presented at the Proceedings of the 2008 symposium on Eye tracking research & applications.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*(3), 323-332.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*(4), 665.

Lange, A., Lehmann, I. J., & Mehrens, W. A. (1967). Using item analysis to improve tests. *Journal of Educational Measurement, 4*(2), 65-68.

Ormrod, J. E. (2012). *Human learning*. Boston: Pearson.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Promes, S. (2013). *McGraw-Hill Specialty Board Review Tintinalli's Emergency Medicine Examination and Board Review 7th edition*: McGraw Hill Professional.

Rosh, A. (2012). *Emergency Medicine PreTest Self-Assessment and Review* (Third ed.): McGraw-Hill Medical.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A., & Pappas, J. (2013). Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department. *Annals of Emergency Medicine, 61*(1), 96-109.

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia-Social and Behavioral Sciences, 29*, 188-197.
Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.
Szulewski, A., Fernando, S. M., Baylis, J., & Howes, D. (2014). Increasing Pupil Size Is Associated with Increasing Cognitive Processing Demands: A Pilot Study Using a Mobile Eye-Tracking Device. *Open Journal of Emergency Medicine, 2014*.
Szulewski, A., & Howes, D. (2014). Combining First-Person Video and Gaze-Tracking in Medical Simulation: A Technical Feasibility Study. *The Scientific World Journal, 2014*.
Tobii. (2012). Tobii Glasses Eye Tracker User Manual. In: Tobii Technology AB.

Chapter 3

Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool

Published as: Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L., & van Merriënboer, J. J. G. (2017). Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 22(4), 951-968.

Abstract

Objective
In general, researchers attempt to quantify cognitive load using physiologic and psychometric measures. Although the construct measured by both of these metrics is thought to represent overall cognitive load, there is a paucity of studies that compares these techniques to one another.

Methods
The authors compared data obtained from one physiologic tool (pupillometry) to one psychometric tool (Paas scale) to explore whether they actually measured the construct of cognitive load as purported. Thirty-two participants with a range of resuscitation medicine experience and expertise completed resuscitation-medicine based multiple-choice-questions as well as arithmetic questions.

Results
Cognitive load, as measured by both tools, was found to be higher for the more difficult questions as well as for questions that were answered incorrectly ($p < 0.001$). The group with the least medical experience had higher cognitive load than both the intermediate and experienced groups when answering domain-specific questions ($p = 0.023$ and $p = 0.003$ respectively for the physiologic tool; $p = 0.006$ and $p < 0.001$ respectively for the psychometric tool). There was a strong positive correlation (Spearman's rho = 0.827, $p < 0.001$ for arithmetic questions; Spearman's rho = 0.606, $p < 0.001$ for medical questions) between the two cognitive load measurement tools.

Conclusion
These findings support the validity argument that both physiologic and psychometric metrics measure the construct of cognitive load.

Introduction

Physician cognitive load is an intrinsic characteristic of work in acute-care medical settings and is known to affect performance (Marx et al., 2013). The nature of work in the emergency department, where physicians are frequently interrupted, treat multiple patients simultaneously and must regularly prioritize decision-making, places considerable demands on their cognitive resources and thus increases the likelihood of making errors (Laxmisan et al., 2007). To greater or lesser degrees, these observations hold true in many domains of medicine where physicians must balance competing priorities while caring for their patients.

From a theoretical perspective, cognitive load is thought to be comprised of three basic elements: intrinsic cognitive load, extraneous cognitive load and germane cognitive load (Young et al., 2014). Intrinsic cognitive load is a function of the complexity of the information to be processed and the expertise of the task performer; while extraneous cognitive load is due to suboptimal information presentation conditions. The sum of intrinsic and extraneous cognitive load is thought to represent the overall cognitive load that can be measured experimentally. Germane cognitive load is thought to refer to the working memory resources dedicated to actively processing intrinsic cognitive load, and thus to learning (Sweller, 2010).

Cognitive load theory (CLT) is a theory of learning based on the optimal design of instructional methods that considers a learner's finite cognitive capacities to apply knowledge and transfer it to new situations (F. Paas et al., 2003). According to CLT, mental processing is limited by the capacity of working memory (De Jong, 2010; Sweller et al., 1998). With the development of domain-specific expertise, those with more experience are thought to be able to chunk related concepts together in elaborated schemas, thus maximizing the efficiency of their working memory (Gegenfurtner et al., 2011; Sweller et al., 1998). In addition to the efficiency afforded by schema creation, individuals are thought to be able to extend domain-specific long-term working memory with experience in a given field despite the traditional supposition that working memory itself is static (Ericsson et al., 1995). In medicine, this is accomplished through the development of retrieval cues between working memory and long-term memory that accelerate memory encoding and decoding.

Richer mental models are created which allow experienced clinicians to more readily recognize when a new clinical scenario may fit with a previously identified pattern (Gegenfurtner & Seppänen, 2013). These same skills allow clinicians to efficiently recognize when a new clinical scenario might not fit a previously identified mental model, thus altering subsequent management decisions (Schubert et al., 2012). Deliberately practicing these cognitive strategies (as well as others) in the context of years of experience allows certain individuals to develop expertise in a domain (Ericsson et al., 2007; Norman, 2005). As a result, a particular task may yield high intrinsic cognitive load for a novice task performer but a much lower intrinsic load for an expert task performer.

For decades, researchers have been interested in measuring cognitive load because it impacts the understanding of expertise development as well as education. It has been shown that measures of cognitive load can reveal important information about CLT beyond traditional performance metrics (F. Paas et al., 2003). The science of cognitive load quantification has been traditionally separated into physiologic measurements and psychometric measurements of this construct (F. Paas et al., 2003). Dual-task performance techniques (which are based on the premise that limited cognitive resources exist that must be distributed between two competing tasks) have gained some popularity in the literature as a means to quantify cognitive load as well (Brunken et al., 2003).

A well-studied method for physiologic measurement of cognitive load is pupillometry. Pupillometry consists of recording a participant's changes in pupil diameter as he/she utilizes cognitive resources for working memory processes. Pupil diameter increases as cognitive load increases as a result of central autonomic nervous system activity. As such, pupillometry is thought to provide an estimate of the intensity of a participant's cognitive load at a given instant in time (Laeng et al., 2012). Numerous studies in various fields have also found pupillometry to be useful to measure cognitive load (Beatty, 1982; Hess, 1965; Hess et al., 1964; Kahneman et al., 1966; Klingner et al., 2008; Klingner et al., 2011; F. Paas et al., 2003; Szulewski et al., 2014).

With respect to resuscitation medicine content and resuscitation medicine expertise, measuring changes in pupil size as a surrogate marker for cognitive load has shown that experienced physicians expend less cognitive load when answering domain-specific multiple choice questions than novices (Szulewski et al., 2015). It is postulated that experts' lower

level of cognitive load in a testing environment is related, in part, to their expertise in authentic clinical situations (like work in an emergency department) and their expanded long-term working memory.

In addition to physiologic measures of cognitive load, psychometric scales that measure subjective cognitive load are widely used in the literature. One such example is the nine-point mental effort scale developed by Paas (F. G. Paas, 1992). This scale has been widely used in the literature and has been shown to be a reliable and valid measure of overall cognitive load (Ayres, 2006; F. Paas et al., 2003). A copy of this scale is included in Appendix 1.

Though both physiologic and psychometric measures attempt to quantify cognitive load, there is no accepted gold-standard for cognitive load measurement. Some authors have questioned whether data derived from psychometric surveys might actually give information about intrinsic cognitive load, as opposed to overall cognitive load, as has been traditionally assumed (Naismith, Cheung, et al., 2015). Others have brought into question whether construct validity truly exists and if another variable, like stress, may actually be the one being measured using these techniques. A recent systematic review on cognitive load measures concluded that the quality of evidence for cognitive load measurement is low and that multiple quantification techniques should be used together in future studies to address this issue (Naismith & Cavalcanti, 2015). In short, consensus about the validity of these measures does not fully exist. Moreover, there is a paucity of studies that compares physiologic and psychometric tools head-to-head. Without more evidence, it would be premature to conclude that they are reliably measuring the same construct and that this construct is, in fact, cognitive load. This study attempts to bridge this gap, by providing evidence of validity using aspects of Cook's review of the Messick validity framework as a guide (Cook et al., 2006). This framework suggests that evidence to support validity of an instrument should be based on information from five sources (content, response process, internal structure, relations to other variables and consequences).

The objective of this experiment was to investigate the relationship of measured cognitive load as determined by (1) an analysis of changing pupil size and (2) responses to a subjective psychometric mental effort questionnaire. These experiments were carried out in

participants with varying levels of resuscitation medicine experience as they performed a resuscitation medicine test.

Methods

Experimental setting

Participant cognitive load was measured by both physiologic and psychometric measures as participants with varying levels of resuscitation medicine experience answered a multiple-choice question (MCQ) test presented to them on a computer monitor. A research assistant who was not involved in data analysis or experimental design conducted the experiment with each participant.

Participants

A convenience sample of 32 participants was recruited between September and November of 2014. Participants were grouped according to their experience in resuscitation medicine. The novice group comprised 13 undergraduate medical students in their first two years of medical school. The intermediate group consisted of 9 senior residents (fourth or fifth year residents enrolled in emergency medicine and other resuscitation-based fields) as well as emergency medicine attending physicians in their first years of practice. The experienced group of participants included 10 attending physicians with more than ten years of clinical experience in fields related to emergency and resuscitation medicine.

The mean age of the 32 participants was 34.1 (SD = 10.8) years. The mean experience level, defined as number of years since starting medical school for all participants, was 10.1 (SD = 10.2). Participant mean age was 24.3 (SD = 1.8), 33.1 (SD = 2.6), and 47.7 (SD = 7.2) years for the novice, intermediate and experienced groups, respectively. Female participants made up 6 of the 13 novice subjects, 1 of 9 intermediate participants and 2 of 10 experienced participants. It had been a mean of 0.8 (SD = 0.4), 9.3 (SD = 2.1), and 22.9 (SD = 7.2) years since the start of medical school for the novice, intermediate and experienced groups respectively.

The rationale for this division of participants was to provide evidence of "*relations with other variables"* for the validity argument of the testing instruments used. *Relations with other variables evidence is* thought to bolster the validity argument when the results from subgroups based on training status vary as expected. The authors hypothesized that the novice group would have a relatively low content-knowledge, but a high test-taking ability as a result of their temporal proximity to similar MCQ testing. The intermediate group was expected to have both a high content-knowledge as well as test-taking ability. In contrast, the experienced group was hypothesized to have a high content-knowledge but a relatively lower test-taking ability because of the increased time elapsed from their own written MCQ examinations.

Thirty-five participants were contacted by email by one of the authors to take part in the study; they did not receive an incentive to participate. Eligibility was determined based upon known training/experience level. One potential participant from the intermediate group and two potential participants from the experienced group declined to participate. All novices approached agreed to participate. The Research Ethics Board at Queen's University provided approval for this study (SMED–115-13; extension of file #6010680).

Tools used

Prior to each individual session, participants were fitted with the Tobii© Glasses Eye Tracker (Tobii Technologies, Danderyd, Sweden) and the device was calibrated as per manufacturer recommendations. During this calibration, each participant's pupil size at baseline was determined. The equipment subsequently calculated the dynamic monocular pupil size as a percentage of baseline at a rate of 30 Hz throughout the experimental session.

After answering each question, participants were prompted to rate their mental effort using the psychometric mental effort questionnaire developed by F. Paas (1992). See Appendix 1 for details. Participants verbalized their responses to the questions and surveys; these audio data were recorded synchronously and analyzed later by a research assistant blinded to the pupillometry data.

These two tools were utilized in the current experiment in an effort to provide validity evidence based on "*relations to other variables"*. This source of evidence for validity is based on the idea that if the tools are measuring the same construct, then there should be a correlation between their scores (Cook et al., 2006).

Furthermore, the Paas scale was used in an effort to see whether the pupillometry data was measuring what it was purported to measure based on the thought process of the participants. If the actions (pupillometry data in this case) fit with the thought process of participants, this would provide evidence validity based on "*response process"* (Cook et al., 2006). Because the Paas scale asks participants to rate their level of investment of cognitive resources and pupillometry is supposed to quantify the investment of cognitive resources, it was theorized that a correlation between the two tools would provide some of this "*response process"* evidence.

Instrument validity and reliability

Using various measuring devices, researchers have been using pupillary measurements as a surrogate marker for cognitive load and have validated its use in numerous experimental realms. Beatty (1982) found that digit and linguistic tasks of increasing complexity caused pupillary size to increase to greater degrees. In an early experiment, Hess (1965) showed that pupil size increased when arithmetic problems were presented to participants, peaked when the answer was given and then dropped off again. Further, higher peak pupil dilation has been shown to be associated with increasing task difficulty (Klingner et al., 2011). Szulewski et al. (2014) were able to replicate these findings using arithmetic questions with newer mobile eye-tracking technology. This new technology has also been successfully utilized in cognitive load measurement experiments in medical testing where questions posed to participants were shown to affect pupil size in predictable ways based on question and participant characteristics (Szulewski et al., 2015). Other groups of researchers have also found consistent results using the mental effort scale developed by Paas found in Appendix 1 (Ayres, 2006; Tuovinen et al., 2004). For example, in a group of high school students solving algebra problems, Ayres (2006) showed that the Paas scale provided a cognitive load rating that was reliable and correlated highly with errors, as expected.

Experimental design

After calibration of the eye-tracking device, each participant sat at a distance of one metre from a computer monitor on which questions were displayed. Ambient light and screen brightness were standardized and participants were asked to continue looking at the screen throughout the experiment and to verbalize their responses to the questions.

Each participant encountered the same questions in the same order. Four arithmetic questions were interspersed among twelve resuscitation-based medical MCQ's. Questions were classified *a priori* into "difficult" and "easy" questions based on their origin (medical student handbook vs. specialty board examination preparation material) as well as the authors' judgement in an effort to provide evidence from a "*relations to other variables"* source. A black circle was presented on the monitor between questions to re-establish a pre-question pupil diameter baseline for each question and each participant.

After each question, participants were prompted to rate their mental effort using the mental effort scale in Appendix 1 (Paas scale).

Data analysis

In general, when a participant reads a question or is presented with a problem to solve, his/her pupil diameter increases steadily until he/she provides an answer, at which point the pupil diameter decreases again (Kahneman et al., 1966). Previous studies utilizing pupillometry as a physiologic measure of cognitive load have concluded that measuring cognitive load accurately requires an analysis of both the magnitude of the change in pupil size as well as the duration of time between question presentation and verbalization of a response (ie. the time that a participant is thinking about an answer) (Szulewski et al., 2015).

To combine these two parameters into one measure we calculated the area under the curve (above baseline) of the change in pupil size (expressed as a percentage of baseline pupil diameter at time of calibration) versus time from question start to verbalization of an answer [this is referred to as *pupillary change index* (PCI) throughout this manuscript and expressed in units of % seconds]. The size of this value was hypothesized to represent the participant's overall cognitive load for a given question. The determination of this value was

accomplished by manual graphical analysis of the raw data to obtain a quantitative measure for each participant and each question. In order to account for possible baseline drift or residual cognitive load from a previous question, the baseline value for pupil size was recalibrated for each question by averaging the raw data just before each question was presented (corresponding to the time that each participant was focusing on a black circle presented between each question). See Appendix 2 for a visual representation of one example of the raw pupillometric data. The manual determination of the PCI was labour-intensive (about 15 minutes for each PCI); accordingly, we decided to focus this analysis on the first half of both the arithmetic and medical questions for all participants.

Peak pupil size was determined for all questions and all participants. To aid in this analysis, a procedure in Visual Basic was implemented to clean and analyze the pupillometry data (available from the authors on request). To reduce artefactual (e.g. blinking) and missing data, the 30Hz raw data was smoothed by replacing values that were blank or deviated by more than 10% absolute from the previous value with the rolling average of the previous 1/6 of a second.

The psychometric survey responses and the physiologic pupil data were then compared by question type, level of participant experience, correctness and level of question difficulty.

Correlational analyses were performed using parametric statistics (Pearson correlation) as well as non-parametric statistics (Spearman's rho) as not all data were normally distributed. Other analyses were made by Pearson chi-square, student's t-test (non-parametric Mann-Whitney U), ANOVA (non-parametric Kruskal-Wallis), and post-hoc Tukey analysis. IBM SPSS Statistics 21 was used for all analyses. Correlation effect sizes were designated *a priori* as weak (0.10 – 0.29), moderate (0.30 – 0.49) and strong (≥ 0.50) (Cohen, 1988). Differences were considered to be significant at a level of $p < 0.05$.

Results

i. Correlation between PCI and Paas scale

Overall, the PCI (a physiologic measure of cognitive load) correlated well with the Paas scale (a psychometric measure of cognitive load).

For the arithmetic questions, parametric analysis revealed a Pearson's correlation coefficient of 0.675 ($p < 0.001$), which indicates a strong positive relationship. For the medical questions, parametric analysis revealed a Pearson's correlation coefficient of 0.542 ($p < 0.001$), which also indicates a strong positive relationship. Non-parametric correlational analyses were also performed in order to confirm these values [Spearman's rho for arithmetic questions was 0.827 ($p<0.001$); Spearman's rho for medical questions was 0.606 ($p<0.001$)]. See Figure 1 for a scatterplot of PCI values plotted against Paas scale result values. These strong correlations persisted when analyzed by participant level of experience (Table 1).

Table 1: Parametric (Pearson's r) and non-parametric (Spearman's ρ) correlation coefficients of pupillary change index versus Paas scale response for the arithmetic and medical questions, broken down by sub-group.

	Arithmetic questions		Medical questions	
	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
Novice	0.644 ($p = 0.003$)	0.716 ($p = 0.001$)	0.422 ($p = 0.001$)	0.501 ($p < 0.001$)
Intermediate	0.807 ($p < 0.001$)	0.834 ($p < 0.001$)	0.561 ($p < 0.001$)	0.607 ($p < 0.001$)
Experienced	0.636 ($p = 0.003$)	0.851 ($p < 0.001$)	0.639 ($p < 0.001$)	0.611 ($p < 0.001$)
Overall	0.675 ($p < 0.001$)	0.827 ($p < 0.001$)	0.542 ($p < 0.001$)	0.606 ($p < 0.001$)

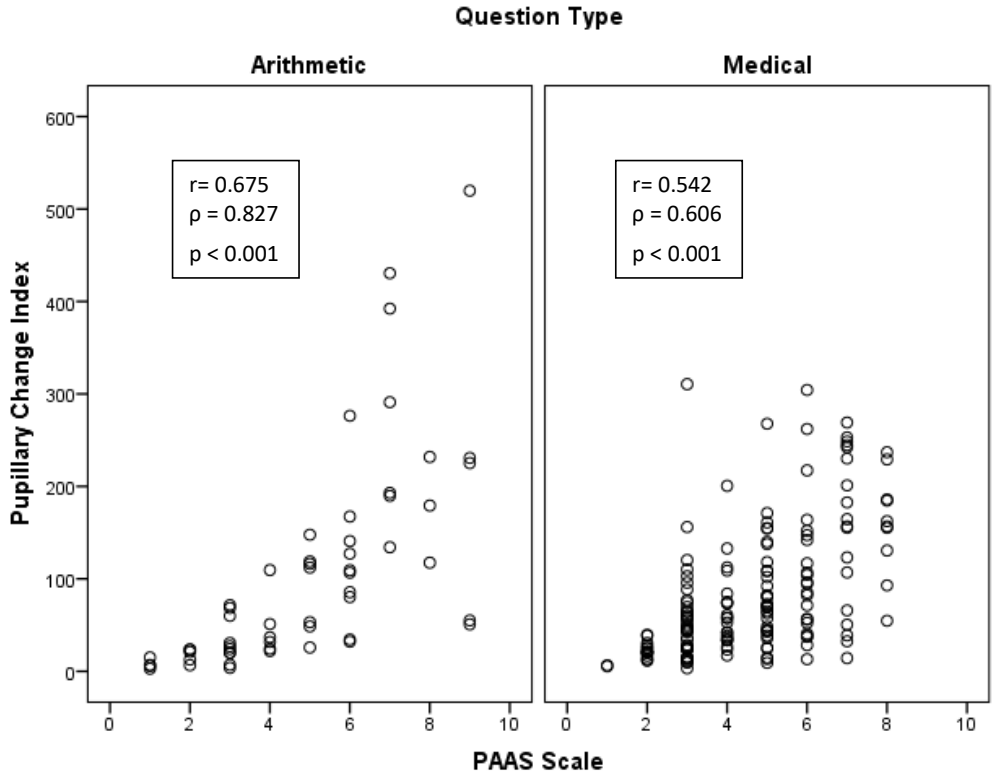


Figure 1: Graphical representation of the correlation between pupillary change index (in % seconds) plotted against Paas scale response for all participants (novice, intermediate and experienced). (r = Pearson's r; ρ = Spearman's rho)

ii. Performance based on training subgroup

Table 2 provides a summary of performance by experimental group as well as question type. There was no significant difference in performance on the arithmetic questions between the novice group and the intermediate group, the novice group and the expert group and the intermediate group and the expert group ($p = 0.858$, $p = 0.410$, and $p = 0.555$ respectively). For the medical questions, the novice group performed significantly worse than both the experienced and intermediate groups ($p < 0.001$). Though both the intermediate and experienced groups were fairly accurate, the intermediate group significantly outperformed the experienced group ($p = 0.044$).

Table 2: Mean (95% confidence interval) proportion of questions answered correctly by subgroup.

	Arithmetic questions	Medical questions
Novice	73% (59 – 93%)	32% (25 – 40%)
Intermediate	74% (58 – 86%)	89% (82 – 94%)
Experienced	80% (65 – 90%)	79% (71 – 85%)

iii. Additional evidence for PCI

a. Variation across training subgroup and question type

Arithmetic questions:

The PCI results for novices were higher than for experienced participants when answering arithmetic questions ($p = 0.005$). There was no significant difference between the intermediate and experienced groups ($p = 0.443$). In addition, there was no significant difference between the intermediate group and the novice group ($p = 0.128$).

Medical questions:

The PCI values of novices were significantly higher than for experienced participants for the medical questions ($p = 0.003$). This same pattern was observed when comparing the PCI result of novices to intermediate participants ($p = 0.023$). There was no significant difference when the PCI values of intermediate participants were compared to experienced participants ($p = 0.851$).

b. Variation across item difficulty, correctness and question type

Overall, difficult questions were associated with a substantially higher mean PCI than easy questions (123.40 vs. 45.14; $p < 0.001$). Similarly, incorrectly answered questions were associated with a higher PCI compared to correctly answered questions (152.51 vs. 66.04; $p < 0.001$). There was no significant difference in PCI when the arithmetic questions were compared to the medical questions (101.52 vs. 89.69; $p = 0.42$). See Table 3 for details.

Table 3: Effect of question difficulty, correctness and type on mean pupillary change index and mean Paas score for all subgroups. PCI = pupillary change index; SD = standard deviation

	PCI mean (SD) in % seconds	Paas score mean (SD)
Difficult questions	123.40 (108.26)	5.40 (1.80)
Easy questions	45.14 (37.52)	3.62 (1.62)
Easy vs difficult	$p < 0.001$	$p < 0.001$
Correct responses	66.04 (69.46)	4.22 (1.80)
Incorrect responses	152.51 (117.82)	5.89 (1.74)
Correct vs incorrect	$p < 0.001$	$p < 0.001$
Arithmetic questions	101.52 (111.22)	4.89 (2.32)
Medical questions	89.69 (89.61)	4.66 (1.80)
Arithmetic vs medical	$p = 0.42$	$p = 0.43$

c. Internal structure evidence

Of 232 possible data points, 10 were missing in the PCI data set because of poor pupillary size output quality in the raw data. Values were not normally distributed and were skewed toward smaller PCIs. There were a limited number of outliers, all on the high PCI side. See Appendix 3 for additional details.

iv. Additional evidence for Paas scale

a. Variation across training subgroup and question type

Arithmetic questions:

Similar to the PCI findings, the Paas scale results for novices were higher than for experienced participants when answering arithmetic questions ($p = 0.040$). There was also no significant difference between the intermediate and experienced groups ($p = 0.549$). In addition, there was no significant difference between the intermediate group and the novice group ($p = 0.365$).

Medical questions:

In keeping with the PCI results, the Paas scale values of novices were significantly higher than for experienced participants for the medical questions ($p < 0.001$). This same pattern

was observed when comparing the Paas scale result of novices to intermediate participants ($p = 0.006$). There was no significant difference when the Paas scale values of intermediate participants were compared to experienced participants ($p = 0.279$).

b. Variation across item difficulty, correctness and question type

In keeping with the PCI results, difficult questions were associated with a higher Paas scale rating than easy questions (5.40 vs. 3.62; $p < 0.001$). Similarly, incorrectly answered questions were associated with a higher Paas scale rating than correctly answered questions (5.89 vs. 4.22; $p < 0.001$). There was no significant difference in Paas scale rating when the arithmetic questions were compared to the medical questions (4.89 vs. 4.66; $p = 0.43$). See Table 3 for details.

c. Internal structure evidence

All 232 possible data points for the Paas scale were collected, with no missing values. The data were normally distributed, with no outliers. See Appendix 3 for additional details.

v. Peak pupil size analysis

Analysis of peak pupil size data for the arithmetic questions revealed no significant differences between any of the subgroups (novice and intermediate $p = 0.15$; novice and experienced $p = 0.08$; intermediate and experienced $p = 0.97$). See Table 4 for descriptive statistics.

Analysis of peak pupil size for the medical questions revealed a significant difference in peak pupil size between the novice group and the experienced group ($p = 0.002$). There was no significant difference between peak pupil size in the novice and intermediate groups ($p = 0.38$) or the intermediate and experienced groups ($p = 0.10$). See Table 4 for descriptive statistics.

Table 4: Pupillary change index, Paas score and peak pupillary size by subgroup and question type. PCI = pupillary change index; SD = standard deviation

	PCI mean (SD) in % seconds	Paas score mean (SD)	Peak pupillary size mean in % (SD)
Novice (arithmetic)	160.70 (152.10)	5.84 (2.31)	112.55 (SD = 14.03)
Intermediate (arithmetic)	93.60 (80.02)	4.82 (2.13)	106.66 (SD = 16.56)
Experienced (arithmetic)	52.44 (52.09)	4.05 (2.24)	105.91 (SD = 8.09)
Novice (medical)	121.39 (113.32)	5.46 (1.75)	101.76 (10.89)
Intermediate (medical)	76.88 (67.60)	4.47 (1.83)	100.05 (6.69)
Experienced (medical)	67.73 (67.93)	3.98 (1.50)	97.33 (10.46)

Discussion

In this study, we compared two methods of cognitive load quantification – a physiologic measure (pupillometry with time to response) and a psychometric measure (Paas scale). In addition, we examined the relationship between cognitive load measures and experience levels of the participants. The goal was to add to the body of literature that supports the validity of using these techniques to measure cognitive load, using Cook's review of the Messick framework as a guide.

A direct comparison of the PCI and Paas scale values revealed a strong positive correlation. Further analysis revealed that this correlation was consistent within subgroups of experience level. This suggests that both the PCI and the Paas scale measured the same construct to some degree. Given previous work in this field as well as the confirmatory results from this experiment, it is likely that this construct is indeed overall cognitive load, which is thought to represent the sum of intrinsic and extraneous cognitive load (Sweller, 2010). We suggest that the correlation found between the two instruments studied provides evidence for validity with respect to Cook's description of "*relations to other variables"* as both variables are commonly used in the literature to quantify cognitive load.

We also found that difficult arithmetic and difficult medical questions resulted in increased cognitive load compared to easier questions, both when analyzed using the PCI as well as the Paas questionnaire. Using both pupillometry and psychometric analysis, we found that incorrectly answered questions caused participants to experience more cognitive load than questions they answered correctly. Finally, no difference was found when questions were divided into arithmetic and medical subgroups. Pupillometry and psychometric results followed the same patterns for all of these analyses. These findings are reassuring as they are in keeping with results from previous studies that show similar trends (Szulewski et al., 2015). In sum, difficult and incorrectly answered questions caused participants to experience greater cognitive load, regardless of question type. These results are expected and bolster the validity argument from a "*relations to other variables"* source.

The PCI and Paas scale data for the arithmetic questions showed that novices experienced higher cognitive load compared to experts when answering questions despite no significant difference in performance between the groups. Of note, there was no significant difference between the peak pupil size between the groups when answering the arithmetic questions. The lack of significant difference in performance on the arithmetic questions makes sense as participants were divided into groups based on resuscitation medicine experience, not arithmetic experience. The lower cognitive load experienced by the experienced group compared to the novice group for the arithmetic questions may be related to age. It is possible that these older participants were either better at arithmetic or that physiologic changes of ageing were responsible for a less responsive pupil. The lack of difference in the peak pupil size analysis for the arithmetic questions makes this latter point less likely. Importantly, both PCI and Paas scale data were consistent with one another when considering the arithmetic questions.

An analysis of the PCI and Paas scale data for the medical questions suggested that novices had higher cognitive load than both the intermediate and experienced groups. Further, no significant difference was found in cognitive load of the intermediate group compared to the expert group for these medical questions. Analogous to these trends, novices performed significantly worse than both the experienced and intermediate groups. Conceptually, these findings may be explained by the fact that novices have relative inexperience with

resuscitation medicine content material, leading them both to perform more poorly and to experience higher cognitive load when attempting to find solutions to problems. The lack of difference in cognitive load of medical questions on physicians in the intermediate versus the experienced group is not surprising given that both these groups are well versed in the content material comprised in the experimental test. The intermediate group's significantly increased score over the experienced group is likely related to this group's relative proximity to their specialty examinations. Finally, the medical question analysis revealed that novices experienced a significantly increased peak pupil size compared to the experienced group. This difference was not present when answering arithmetic questions.

Together, these observations emphasize that cognitive load measurement by both physiologic and psychometric tools acts in a way that is expected and explainable across groups of physicians with varying levels of experience as well as between question types (domain-specific medical questions vs arithmetic questions). These patterns provide some evidence of what Cook would call validity from a "*relations to other variables"* source. That is, the observed patterns in the data varied across groups of participants with different training status as well as with question type, as expected.

The data distribution presented in Appendix 3 provides some validity evidence from an *internal structure* source. The analysis captured all Paas responses and missed less than 5% of PCI values (due to poor data quality). No outliers were identified in the Paas responses and a relatively small number of outliers were identified in the PCI analysis.

Finally, this study provides some indirect evidence of *response process* as a source of validity. Response process, as a source of validity, exists when the actions and thought processes of participants align with the intended measured construct. In this study, though participants were not asked to specifically describe their thought processes (as may have been done with a think-aloud protocol), the Paas survey asked participants to rate investment of cognitive resources, which is what the pupillometry metrics were designed to measure.

Although this paper provides some evidence to support that both physiologic and psychometric measures of cognitive load quantify cognitive load as a construct to some

degree, each has its own strengths and limitations. Psychometric scales are easy to use and cheap to implement, whereas pupillometry is expensive and not practical for routine use in the real world (although this will likely change as the cost of the technology decreases). On the other hand, psychometric scales are prone to participant manipulation and only provide a single cognitive load measurement. Conversely, pupillometry has the advantage of being objective, difficult to manipulate and provides real-time data throughout the peaks and troughs of cognitive effort during the completion of a task.

Importantly, we found the pupillometric data analysis to be time-consuming and difficult to automate with programmed code in this experiment as a result of changing baseline pupil sizes between questions. Changes to the experimental design by minimizing interruptions and more consistently defining pupil baseline (possibly with a standardized cognitive task), may facilitate data extraction in the future. Until this is further addressed, it would be a reasonable choice to use the Paas questionnaire as a means to determine cognitive load in a test-taking setting, especially if a general understanding (as opposed to a real-time and detailed) assessment is sought. The peak pupil size variable, which is much easier to extract, is another available option if a physiologic variable is desired. Ultimately, a real-time physiologic measure obtained unobtrusively, like pupillometry, has powerful implications for the delivery and assessment of learning within CLT.

Our study has certain limitations. To begin, we were unable to control for participant age between groups. This is unavoidable given the inherent nature of experience, but it does raise questions about the possible confounding effects of age on both pupillometry analysis as well as questionnaire responses. Although we recognize that there might be an effect, we believe it to be small given that pupillary size and psychometric responses varied in the same direction throughout this study. Further, the fact that there was no statistical difference in peak pupil size in the arithmetic questions between groups (but there was for the medical questions) raises doubts that the PCI findings are solely caused by physiological pupillary changes related to ageing.

Secondly, we used known-groups comparisons in part of our analysis of validity from a "*relations to other variables"* source. Though necessary, this type of comparison (on its own) is known to be non-specific and inconclusive because of possible confounding effects (Cook,

2015). The analysis of training-relevant (medical) and training-irrelevant (arithmetic) questions strengthens this argument, but the possibility of confounding remains.

In addition, we did not distinguish between the types of cognitive load (intrinsic vs. extraneous vs. germane) and instead chose to focus on total measurable cognitive load. Although this strategy was necessary in order to investigate the research question, a deeper analysis using an instrument designed to differentiate types of cognitive load like the one proposed by Leppink et al. (2014) may have been beneficial – especially given findings in recent literature that have brought into question whether psychometric data may actually be measuring intrinsic, as opposed to total, cognitive load in certain settings (Naismith, Cheung, et al., 2015).

The fact that two potential participants from the experienced group and one potential participant from the intermediate group declined to participate in the study could lead to selection bias. In addition, the low proportion of female participants in this study has the potential to skew the results; however, we do not know of any literature that supports a difference in pupillary responses based on sex.

Finally, although we found a strong positive relationship between our two measures and thus were able to conclude that both sets of data are likely measuring the same construct to a large degree, we cannot absolutely confirm that this construct is indeed overall cognitive load, as opposed to another, related concept. Despite this, we feel that our data triangulates the available evidence and strengthens the argument that overall cognitive load is, in large part, the construct that is being measured by these techniques. From the perspective of Cook's review of Messick's framework of validity, although we provide some evidence of *internal structure*, *response process* and *relations to other variables*, this study does not address *content* and *consequences* as sources for validity evidence.

Our work focused on cognitive load measurement in a tightly controlled (not true-to-clinical-life) MCQ environment. Future studies that measure cognitive load in real (or at least simulated) medical emergencies could provide more accurate insights into medical decision-making and *in situ* cognitive load. In addition, pupillometry has the potential to complement research in visual expertise (Gegenfurtner et al., in press). It is well recognized

that novices and experts have different visual gaze patterns in a variety of professional domains (Gegenfurtner et al., 2011; Gegenfurtner, Siewiorek, et al., 2013; Gegenfurtner et al., 2016; Kok et al., 2012). Measuring cognitive load via pupillometry while analyzing participants' visual patterns as they perform selected tasks could provide further insights into the cognitive process and how it changes with expertise. This type of experiment would be fairly easy to perform as the eye-tracking tool used in this study can track gaze behaviours and record pupillometry data simultaneously.

The currently accepted understanding of validity places an emphasis on construct validity as the "whole" of the validity argument (Downing, 2003). Based on the results of this study, we have been able to provide further evidence of construct validity that the PCI and the psychometric Paas scale are indeed reasonable surrogate markers of cognitive load. Researchers and educators can have increased confidence using either measure depending on the context and the purpose of their study, as they both appear to measure the same construct.

Conclusion

Comparing the measurement of a construct thought to represent cognitive load on medical professionals (both with a pupillometry-based physiologic tool as well as a psychometric survey) reveals a strong positive correlation between the two techniques as well as expected patterns based on question type, difficulty, correctness of answers and training status. This provides evidence that the construct being measured in both cases is related. Overall, the results support the validity of using data obtained using either technique as a surrogate for cognitive load. Further study into the subtypes of cognitive load in medical testing environments as well as cognitive load measurement in real-life clinical scenarios has the potential to provide new insights into the clinical decision-making process.

Acknowledgements

The authors would like to thank Wilma Hopman for assistance with statistical analysis, Bence Linder for development and implementation of the algorithm to smooth the raw pupillometry data and to calculate peak pupillary size, as well as Dr. Jimmie Leppink for advice about experimental design. The authors would also like to acknowledge the Kingston Resuscitation Institute for providing access to the pupillometry device and research assistants.

Appendix 1

Psychometric survey used in the study, adapted from Paas (1992).

Please choose the category (1, 2, 3, 4, 5, 6, 7, 8, or 9) that applies to you:

In the exercise that just finished, I invested:

1. very, very low mental effort

2. very low mental effort

3. low mental effort

4. rather low mental effort

5. neither low nor high mental effort

6. rather high mental effort

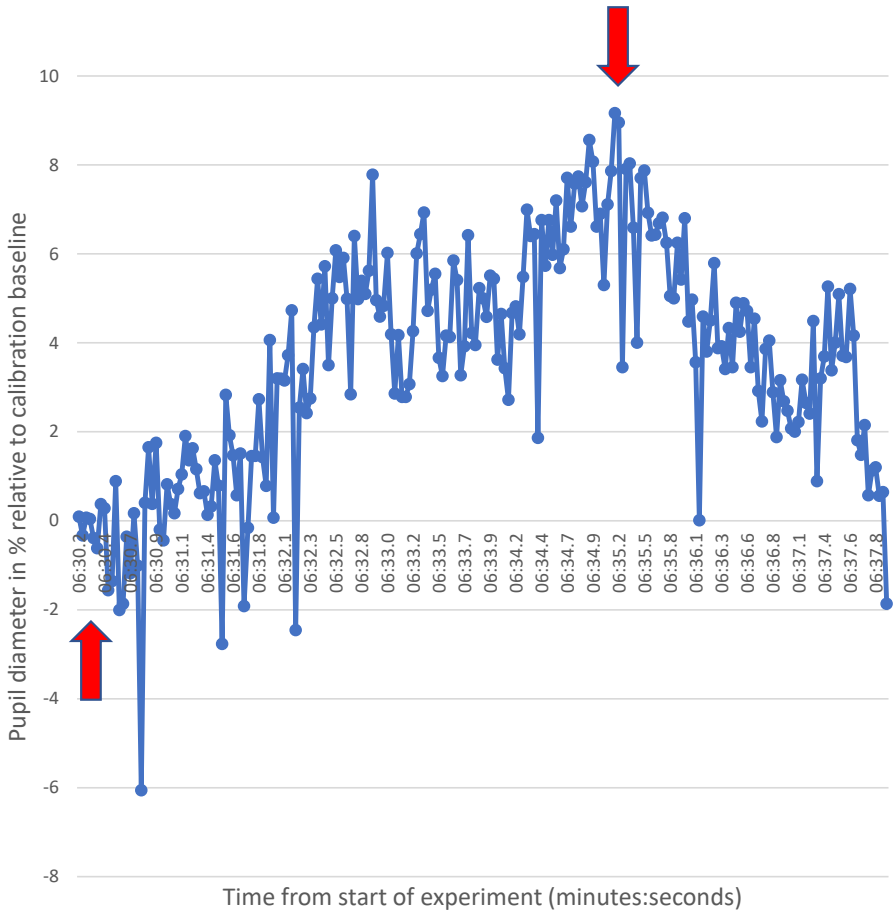
7. high mental effort

8. very high mental effort

9. very, very high mental effort

Appendix 2

Example of raw pupillometry data obtained from one experienced participant for one medical question. The first arrow represents the time the question appeared on the screen. The second arrow represents the point at which the participant verbalized his answer. As the participant experiences increasing cognitive load during the thought process, the pupil diameter increases in size. When the participant verbalizes the answer to the question, pupil size decreases again. The quantitative cognitive load measurement used in the pupillometry arm of this study can be conceptualized as the area under the curve between these two arrows (referred to as pupillary change index in this manuscript).

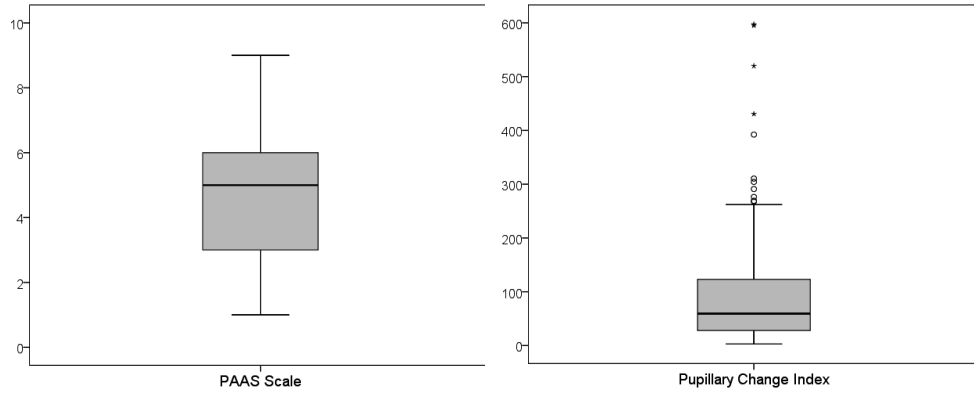


Appendix 3

Data distribution for Paas and Pupillary Change Index scales:

	Paas	PCI
Valid data points	232	222
Missing data points	0	10
Minimum value	1	2.9
Maximum value	9	597.2
Mean	4.7	92.7
Standard deviation	1.9	95.5
25th percentile	3	27.7
Median	5	59.4
75th percentile	6	124.2

Boxplots of the data distribution of the Paas and Pupillary Change Index scales showing outliers:



References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*(5), 389-400.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276-292.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53-61.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd edn. Hillsdale, New Jersey: L: Erlbaum.

Cook, D. A. (2015). Much ado about differences: why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education, 20*(3), 829-834.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine, 119*(2), 166. e167-166. e116.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*(2), 105-134.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological review, 102*(2), 211.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7/8), 114.

Gegenfurtner, A., Kok, E., Van Geel, K., De Bruin, A., Jarodzka, H., Szulewski, A., & Van Merriënboer, J. J. G. (in press). The challenges of studying visual expertise in medical image diagnosis. *Medical Education*.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552.

Gegenfurtner, A., & Seppänen, M. (2013). Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations. *Computers & Education, 63*, 393-403.

Gegenfurtner, A., Siewiorek, A., Lehtinen, E., & Säljö, R. (2013). Assessing the quality of expertise differences in the comprehension of medical visualizations. *Vocations and Learning, 6*(1), 37-54.

Gegenfurtner, A., & Szulewski, A. (2016). Visual expertise and the Quiet Eye in sports – comment on Vickers. *Current Issues in Sport Science, 1*(1).

Hess, E. H. (1965). Attitude and pupil size. *Scientific american*.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583-1585.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). *Measuring the task-evoked pupillary response with a remote eye tracker.* Paper presented at the Proceedings of the 2008 symposium on Eye tracking research & applications.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*(3), 323-332.

Kok, E. M., Bruin, A. B., Robben, S. G., & Merriënboer, J. J. (2012). Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. *Applied Cognitive Psychology, 26*(6), 854-862.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry A Window to the Preconscious? *Perspectives on Psychological Science, 7*(1), 18-27.

Laxmisan, A., Hakimzada, F., Sayan, O. R., Green, R. A., Zhang, J., & Patel, V. L. (2007). The multitasking clinician: decision-making and cognitive demand during and after team handoffs in emergency care. *International Journal of Medical Informatics, 76*(11), 801-811.

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P., & van Merriënboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32-42.

Marx, J., Walls, R., & Hockberger, R. (2013). *Rosen's Emergency Medicine-Concepts and Clinical Practice*: Elsevier Health Sciences.

Naismith, L. M., & Cavalcanti, R. B. (2015). Validity of Cognitive Load Measures in Simulation-Based Training: A Systematic Review. *Academic Medicine, 90*(11), S24-S35.

Naismith, L. M., Cheung, J. J., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education, 49*(8), 805-814.

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Medical Education, 39*(4), 418-427.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A., & Pappas, J. (2013). Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department. *Annals of Emergency Medicine*, *61*(1), 96-109.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Szulewski, A., Fernando, S. M., Baylis, J., & Howes, D. (2014). Increasing pupil size is associated with increasing cognitive processing demands: A pilot study using a mobile eye-tracking device. *Open Journal of Emergency Medicine*, *2*(01), 8.

Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987.

Tuovinen, J., & Paas, F. (2004). Exploring Multidimensional Approaches to the Efficiency of Instructional Conditions. *Instructional Science, 32*(1-2), 133-152. doi: 10.1023/B:TRUC.0000021813.24669.62

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: Implications for medical education: AMEE guide no. 86. *Medical Teacher, 36*(5), 371-384.

Chapter 4

A New Way to Look at Simulation-Based Assessment: The Relationship Between Gaze-Tracking and Exam Performance

Published as: Szulewski, A., Egan, R., Gegenfurtner, A., Howes, D., Dashi, G., McGraw, N. C., Hall, A. K., Dagnone J. D., & van Merriënboer, J. J. G. (2018). A new way to look at simulation-based assessment: the relationship between gaze-tracking and exam performance. *Canadian Journal of Emergency Medicine*, 1-9.

Abstract

Objective

A key task of the team leader in a medical emergency is effective information gathering. Studying information gathering patterns is readily accomplished with the use of gaze-tracking glasses. This technology was used to generate hypotheses about the relationship between performance scores and expert-hypothesized visual areas of interest in residents across scenarios in simulated medical resuscitation examinations.

Methods

Emergency medicine residents wore gaze-tracking glasses during two simulation-based examinations (n = 29 and 13 respectively). Blinded experts assessed video-recorded performances using a simulation performance assessment tool that has validity evidence in this context. The relationships between gaze patterns and performance scores were analyzed and potential hypotheses generated. Four scenarios were assessed in this study: diabetic ketoacidosis, bradycardia secondary to beta-blocker overdose, ruptured abdominal aortic aneurysm and metabolic acidosis caused by antifreeze ingestion.

Results

Specific gaze patterns were correlated with objective performance. High performers were more likely to fixate on task-relevant stimuli and appropriately ignore task-irrelevant stimuli compared with lower performers. For example, shorter latency to fixation on the vital signs in a case of diabetic ketoacidosis was positively correlated with performance ($r=0.70$, $p < 0.05$). Conversely, total time spent fixating on lab values in a case of ruptured abdominal aortic aneurysm was negatively correlated with performance ($r= -0.50$, $p < 0.05$).

Conclusion

There are differences between the visual patterns of high and low-performing residents. These findings may allow for better characterization of expertise development in resuscitation medicine and provide a framework for future study of visual behaviours in resuscitation cases.

Introduction

Successful management of a medical emergency demands effective crisis resource management (CRM) skills (Kim, Neilipovitz, Cardinal, Chiu, & Clinch, 2006; Watkins, Roberts, Boulet, McEvoy, & Weinger, 2017). Physicians skilled in CRM are usually easily identifiable by their peers, but describing specific behaviours that make them successful in crisis settings is difficult, even for physicians themselves (Szulewski & Howes, 2014). Part of the reason is that high-level CRM skills are automatized over many years as a result of deliberate practice and become second-nature (Ericsson, Krampe, & Tesch-Römer, 1993). This tacit knowledge is subsequently difficult to describe explicitly and teach (Patel, Arocha, & Kaufman, 1999; Szulewski, Brindley, & Van Merrienboer, 2017).

Tacit knowledge poses particular challenges to medical educators who are tasked with preparing and assessing learners in routine as well as emergency medical cases. It follows that the more that is known about expertise development in the management of medical emergencies, such as resuscitation and CRM, the more likely it will be that both can be effectively and efficiently taught and assessed. Moreover, by triangulating novel objective measures of learner performance with traditional assessment, educators can calibrate both modalities and increase the validity of both subjective and objective means of trainee assessment. This will improve the granularity of resuscitation assessment, with decreased risks to patients – a need that will only increase as medical schools shift to competency-based medical education (CBME).

One commonly used approach to teaching CRM skills well suited to integrating CBME is simulation-based training (Gaba, Howard, Fish, Smith, & Sowb, 2001). High-fidelity medical simulation-based training improves learning by providing a forum for feedback, integration of curriculum, and an opportunity for repetitive practice for routine cases as well as variability of practice for non-routine cases (Helle et al., 2011; Issenberg, Mcgaghie, Petrusa, Lee Gordon, & Scalese, 2005; Seppänen & Gegenfurtner, 2012). Formative and summative assessments in simulation are also gaining popularity (Gegenfurtner, Quesada-Pallarès, & Knogler, 2014), especially in postgraduate medical education where assessment tools are being developed and optimized (Bick et al., 2013; Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013; Hall, Dagnone, Lacroix, Pickett, & Klinger, 2015).

It is difficult to glean a comprehensive view of expertise development in resuscitation medicine by traditional observational research methods. A novel method of studying resuscitation medicine and CRM expertise is with the use of mobile gaze-tracking technology. Gaze-tracking data can provide new insights into clinician behaviour and performance and may lead to improved patient safety practices (Henneman, Marquard, Fisher, & Gawlinski, 2017). Mobile gaze-tracking glasses are worn like a pair of eyeglasses that record both a video of a participant's first-person point-of-view as well as track his/her eye movements, superimposing a gaze-indicator on the first-person video which reveals where the participant is looking in real-time. Based on data that shows what a physician "sees", inferences can be made about differences between novices and experts with respect to their visual attention, situational awareness, and ability to manage competing interests (Szulewski, Roth, & Howes, 2015).

Visual expertise research has shown that there are measurable visual differences between novices and experts in numerous domains that can be quantified using gaze-tracking technology (Gegenfurtner, Lehtinen, & Säljö, 2011; Szulewski, Gegenfurtner, Howes, Sivilotti, & van Merriënboer, 2016). Experts are able to selectively ignore irrelevant information, rapidly recognize patterns, and select appropriate diagnostic schemata to fit what they see (Gegenfurtner et al., 2017). With practice, individuals become better, more efficient interpreters of their surroundings. They are better able to prioritize what is important while simultaneously deprioritizing what is not. Analyzing eye-movements in a non-medical field, Haider and Frensch (1999) termed this phenomenon the *information-reduction hypothesis* and Kok, de Bruin, Robben, and van Merriënboer (2012) hypothesize that there may be parallels to be drawn in medicine.

Gaze-tracking technology has also been used to determine the feasibility of studying physician behaviours during simulated medical emergencies and it has been shown that this technique is practical and useful in this setting (Szulewski & Howes, 2014). However, little is known about whether physicians' expertise in the complex environment of resuscitation medicine can be accurately assessed by studying their visual patterns.

Building on this research, this study explored residents' information gathering techniques by analyzing their initial visual fixation patterns in a simulated resuscitation environment. Specifically, we were interested in uncovering particular gaze patterns used by physicians

that are associated with better exam performance. The objectives of this study were to: (1) examine the relationship between initial visual patterns and performance scores, and (2) explore if these patterns varied in different simulation scenarios.

Methods

Study Design and Participants

A convenience sample of Emergency Medicine (EM) residents at one training site were invited to participate in two exams, each comprised of two simulated scenarios.

The first two scenarios (S1: *diabetic ketoacidosis* and S2: *bradycardia secondary to beta-blocker overdose*) were conducted in February 2014; the second two scenarios (S3: *ruptured abdominal aortic aneurysm (AAA)* and *S4: metabolic acidosis caused by antifreeze ingestion*) were conducted in August 2014. Scenarios and the associated assessment tools were developed through an iterative process that consisted of a template-based scenario development tool. At least 3 expert EM physicians reviewed each case for clinical fidelity. A previous study of EM residents, using similar scenarios developed in this manner, demonstrated excellent reliability (as determined through G-Study analysis) (Hall et al., 2015). The study took place during regularly scheduled bi-annual simulation-based objective structured clinical exams (OSCEs). Participants had previously participated in similar OSCEs and were familiar with the assessment environment and format, however the scenarios used in this study were novel to each participant. Participants did not receive an incentive to patriciate. The study was approved by the Research Ethics Board at Queen's University (EMED-162-11).

Experimental Design

Residents were fitted and calibrated with a mobile gaze-tracking device (Tobii Glasses Eye Tracker, Danderyd, Sweden). They then entered the high-fidelity simulation lab as the leader of a team comprised of a registered nurse (RN) actor and a respiratory therapist (RT) actor, after reading a case stem. Simulation rooms were organized to mimic the resuscitation bay in the emergency department. The RN and RT began the scenarios in the same locations in

the room, however they were free to move around the room based on instructions given by the resident. Participants completed two 10-minute scenarios. Audio/video data from three (non gaze-tracking) cameras were also recorded (Kb Port, Allison Park, Pennsylvania). These videos were subsequently used by the blinded external reviewers to score participant performance.

Analysis

The gaze-tracking device recorded first-person audio/video data and gathered information about pupil position and glint location at a rate of 30 Hz. Using these data, a dynamic, gaze indicator was superimposed on the first-person video by computer software *(Tobii Studio Pro)*. This video was used in the analysis to compute the gaze-tracking variables/fixation areas listed in Table 1. See Figures 1 and 2 for static representations of the data generated by the software. Fixations, in this study, refer to instances when the gaze indicator stopped scanning the environment and landed on an area of interest. The frequency and duration of residents' gazes were scored independently by two trained individuals who manually analyzed each first-person video. Human raters were used in this study to count fixations as computer technology that accurately triangulates eye-tracking data with three-dimensional positional data in a resuscitation room with dynamic areas of interest (where individuals and equipment move) is still under development. Dichotomous actions and visual propensities of participants were captured (see Table 1) and Pearson correlation coefficients were used to determine whether associations existed with performance. These areas of interest were defined *a priori* based on their potential clinical relevance based on input from local medical experts. We focused this analysis on the first 60 seconds of each scenario as we were mostly interested in how residents' initial information gathering patterns correlated with performance scores. Furthermore, our previous experience suggests that most relevant visual fixations occur early in simulated scenarios and that there is increased variability and noise in gaze patterns as simulations progress.

Table 1: Measured gaze-tracking fixations areas

Time in first 60 seconds:	Number of visual fixations in the first 60 seconds on the:
• to first vital check • in silence • looking at vitals • looking at RN • looking at lab values • looking at ECG	• Patient • Registered Nurse • Respiratory Therapist • Vitals • Medication List

Figure 1: A third year residents' gaze fixations within the first minute of a simulation. The numbers represent the order in which the resident looked at each area of interest and the size of the circle represents the relative time spent fixating on each of these points.

Figure 2: A third year residents' gaze fixations within the first minute of a simulation. The colours represent a heat map, where red represents more fixations for a longer duration, followed by yellow and green areas, which represent fewer fixations.

Inter-rater reliability between the two individuals tabulating eye-tracking variables and the expert performance assessors were determined by calculating Intraclass Correlation Coefficients (ICC). For each eye-tracking video, an ICC for aggregated fixation times (total fixation time and time until first vital signs check), and for the net number of object fixations in the first 60 seconds was calculated. Two independent undergraduate reviewers used video time stamps, and systematic video pausing to record gaze total gaze duration and frequencies. Reliability was calculated separately for S1 & S2 and S3 & S4 as only five participants were involved in all four scenarios.

Performance of the residents was assessed by external attending emergency physicians who were blinded to participant identity and level of training. The Queen's Simulation Assessment Tool (QSAT) was used for assessment by the physicians after their review of the non-gaze-tracking recordings. The raters had previous experience and training on using the QSAT in this context. The QSAT has been previously shown to discriminate simulation-based OSCE performance with a similar cohort of EM residents within a CRM context (Hall et al., 2015). The QSAT (see Appendix 1) assesses residents on a 5-point Likert scale from *Inferior*

to *Superior* across four categories (Primary assessment, Diagnostic actions, Therapeutic actions, and Communication) as well as Overall Performance.

Factorial ANOVA was conducted to determine if score differentials were only a function of training level, or alternatively, if gaze-tracking provided information on level of experience that extended beyond the level of training. Grouping of residents by experience in scenarios 1&2 and 3&4 differed due to uneven numbers of participants across PGY years. Of note, CCFP (Year 3) residents were considered independent of FRCPC Year 3 residents.

Results

Participants

Thirteen and twenty-nine residents, respectively, participated in the two exams (five completed both exams). One participant declined to consent to gaze-tracking (but still completed the exam). Because of poor data quality and technical errors, results from 3 cases in the first OSCE and 3 cases in the second were excluded. This resulted in a total of 78 cases that were subsequently analyzed. Participants were EM residents enrolled in the College of Family Physicians of Canada (CCFP) and the Royal College of Physicians and Surgeons of Canada (RCPSC) programs (see Table 2) at one EM training site (mean age 30.0, SD 2.91; 48% were female residents).

Table 2: Number of participants by level of training and simulated case for the analyzed data.

Level of Training	Diabetic Ketoacidosis (DKA)	Bradycardia secondary to Beta-Blocker Overdose (BB)	Ruptured Abdominal Aortic Aneurysm (AAA)	Antifreeze Ingestion
1st Year	0	0	5	5
2nd Year	1	2	6	7
3rd Year	2	2	2	2
4th Year	1	1	3	3
5th Year	3	3	1	1
CCFP (3rd year)	4	4	10	10
Total	11	12	27	28

*Note: A total of 6 cases were excluded due to poor data quality and technical errors

Gaze tracking data

Appendix 2 outlines average fixation data for the scenarios. An acceptable average ICC was found for fixation time ratings between the two individuals who analyzed the gaze-tracking videos in S1&2 (ICC = .87, p <.001, 95% CI [.15, .97]), however, an unacceptable ICC was found for S3&4 (ICC = .39, p <.001, 95% CI [-.22, .71]). An acceptable average ICC was found for net number of object fixations (first 60s) for S1&2 (ICC = .67, p < .032, 95% CI [-.03, .90]), and S3&4 (ICC = .82, p < .001, CI 95% [.59, .92]). Fixation time ratings were thought to be unsatisfactory largely due to the difficulty of manually recording the time and numbering gaze manually. The future development of automated tracking processes would be beneficial to the accuracy of tracking.

Performance scores

The calculated ICC was very good between the two expert assessors in S1 (ICC = .85, p <.01, 95% CI [.45, .96]) and S2 (ICC = .87, p <.001, 95% CI [.02, .97]). Only one external blinded assessor reviewed S3 and S4. Assessors scored participants across four categories for a *total* score; in addition to an *overall* impression score. A two-way mixed ICC was calculated for absolute accuracy between raters for *total* and *overall* scores. Findings showed acceptable reliability between *total* and *overall* score for S1 (ICC = .933, p <.001, 95% CI .65, .98]) and S2 (ICC = .987, p <.001, 95% CI [.95, .99]), S3 (ICC = .95, p <.001, 95% CI [.86, .98]) and S4 (ICC = .87, p <.001, 95% CI [.70, .94]). Due to the congruency between *total* and *overall* scores, and between assessors, an average was taken between the percent *total* and *overall* scores for each scenario to create a new *average score*. This final score was used for subsequent analyses (see Appendix 3). After checking for normality, correlations were computed to determine associations between the data points in Table 1 and simulation performance scores.

Relationship between level of training and performance:

In S1&2, resident years were separated into three categories: Years 1-4 (M S1 = .57, S2 = .60); Year 5, (M S1 = .90, S2 = .93); and CCFP (Year 3), (M S1 = .58, S2 = .56). Main effects

showed statistically detectable difference between groups for S1 $F(2, 8) = 5.32$, $p = .03$, $\beta = .67$, and S2 $F(2, 9) = 4.90$, $p = .04$, $\beta = .65$. There were statistically significant differences between Year 5s and CCFP Year 3s ($p = .034$) in S1. No statistically significant differences were found between groups in S2. In S3&4, resident years were separated into three categories: Years 1-2, (M S3 = .64, S4 = .59; Years 3-5 (M S3 = .75, S4 = .63); and CCFP (Year 3), (M S3 = .66, S4 = .57). Main effects showed no statistically detectable difference between residency groupings in S3 $F(2, 25) = 1.65$, $p = .2$, $\beta = .32$, and S4 $F(2, 25) = .14$, $p = .86$, $\beta = .07$. Due to differences in the number of participants, and the spread of residents' level of training between S1&2 and S3&4, groupings could not be consistent. Independent *t*-tests were used to determine ordering effects across scenarios. No effects of ordering were found ($p > .4$). A sample size calculation was not performed as we maximized the number of residents who were available to participate. Instead, we have provided statistical power calculations (β**)**.

Relationship between gaze-tracking and performance:

The observed correlations between gaze-tracking variables and average performance score are found in Tables 3-5. Statistically significant correlations in S1, S3, and S4 indicated that visually fixating on particular people and objects was correlated with performance.

Table 3: Correlations Between Selected Gaze Indicators and Average Score in Scenario 1 (Diabetic Ketoacidosis)

Variables	1	2	3
1. Average Score	-		
2. Number of RT fixations (first 60 seconds)	-.65*	-	
3. Time to first fixation on vital signs	-.70*	.48	-

*p <.05, **p <.01

Table 4: Correlations Between Selected Gaze Indicators and Average Score in Scenario 3 (Ruptured Abdominal Aortic Aneurysm)

Variables	1	2	3	4
1. Average Score	-			
2. Total time fixating on lab values (first 60 seconds)	-.50*	-		
3. Total ECG fixation time (first 60 seconds)	-.39*	.16	-	
4. Number of RN fixations (first 60 seconds)	-.46*	.54**	.117	-

*p <.05, **p <.01

Table 5: Correlations Between Selected Gaze Indicators and Average Score in Scenario 4 (Metabolic Acidosis)

Variables	1	2
1. Average Score	-	
2. Number of medication list fixations	-.48*	-

*p <.05, **p <.01

Discussion

Residents' gaze-tracking patterns were found to be significantly correlated with objective performance in simulation-based resuscitation examinations. Because of the number of participants and the array of data, correlation was used to elucidate patterns and associations. Given that this was a pioneering study into the use of gaze analytics to establish patterns associated with performance, these results are not put forward as definitive findings; rather, they lay the groundwork for future study in this novel area.

In S1 (unstable patient with diabetic ketoacidosis), performance was strongly positively correlated with decreased latency to checking the patient's vital signs – a task-relevant stimulus (Pearson's r of 0.70, $p < 0.05$). The number of times a participant looked at the RT was negatively correlated with performance (Pearson's r of – 0.65, $p < 0.05$). In this case, the RT did not provide any useful clinical information, nor did he have a predefined script. In addition, the patient did not require any advanced airway intervention and did not have any increased work of breathing, and therefore it is plausible that increased views were not only

task-irrelevant, but potentially associated with help-seeking or other attempts at social validation.

In S2 (beta-blocker overdose), there were no statistically detectable correlations identified. This is likely a result of multiple factors. First, the scenario and corresponding assessment tool were unable to discriminate by level of training. Anecdotally, it seems that there was malalignment between features of scenario design and expected trainee behaviours. For example, many senior trainees did not perform transcutaneous pacing, and instead focused immediately on antidotes and medical management of beta-blocker toxicity. Retrospectively, this was not an unreasonable management decision. However, raters were informed that transcutaneous pacing was an important therapy for this patient, and so likely rated these performances with a lower score.

In S3 [patient with ruptured abdominal aortic aneurysm (AAA)], we found that visual fixations on task-irrelevant stimuli like the laboratory values, ECG and nurse were all strongly negatively correlated with objective performance (Pearson's r of –0.50, –0.39, and –0.46, respectively; $p < 0.05$). A physician having made a diagnosis of a ruptured AAA should be primarily focused (after initial stabilization) on transporting the patient to the operating room for life saving surgery. On video review, it was observed that higher performers expedited patient transfer by concentrating on contacting the vascular surgeon, calling the operating room charge nurse, and ensuring the patient was on portable monitors and ready for a rapid transfer. We did not tabulate these areas of interest in the data collection, as these tasks were not predicted to be relevant, *a priori*.

In S4 (patient with metabolic acidosis caused by antifreeze ingestion), participants needed to provide primary resuscitation and appropriately manage a patient with an undifferentiated metabolic acidosis. The number of times a participant fixated on the medication list was correlated with poorer performance (Pearson's r of – 0.48, $p < 0.05$). These low-performing participants tended to assume that a medication overdose was the cause of the patient's condition and overlooked clues that pointed to ingestion of another toxic compound (antifreeze in this case). Those who performed better identified the appropriate agent and spent less time viewing the patient's medication list.

As in clinical practice, each of the scenarios was different from the others and required medical learners to focus their attention on differing stimuli to be successful. Because these visual patterns differ from scenario-to-scenario, it would be difficult for a learner to feign superior visual behaviours without having the requisite knowledge/experience of a high-performing medical learner.

These findings can be considered within the context of the information-reduction hypothesis. This hypothesis submits that as people gain experience and improve their performance, they are better able to quickly identify (and rapidly deprioritize) task-irrelevant stimuli, while appropriately prioritizing task-relevant stimuli (Haider & Frensch, 1996, 1999). Further, improvement in speed and performance on a task are due, in part, to a reduction in the amount of information that an individual processes at the perceptual level. Newer eye-tracking research adds evidence to this theory as domain experts have been found to have more fixations of longer duration on task-relevant information and fewer fixations of shorter duration on task-irrelevant information (Gegenfurtner et al., 2011).

The decreased ability to appropriately shift visual attention away from task-irrelevant stimuli may be related (at its extreme) to the concept of *helmet fire.* In *helmet fire,* an overload of stimuli overwhelms the working memory resources of a (usually) more inexperienced physician when tasked with the management of a sick patient under high stress conditions. *Helmet fire* decreases task performance and is observed in many novices in both simulation and clinical practice when managing a situation beyond their comfort-level.

In our study, candidates were not specifically presented with task-irrelevant information. Task-irrelevant stimuli could be seen by some as distractors that may artificially inflate difficulty. The authors argue that, in the future, adding *realistic* distractors may increase the opportunity to observe discriminating behaviours that are indicative of expertise.

It may have been expected that we would have identified more positive correlations with task-relevant areas, as well as negative correlations with task-irrelevant areas. A possible explanation was found on video review. While expert avoidance of task-irrelevant stimuli is striking, consistency in viewing task-relevant data is much less obvious and systematic.

Experts may take different paths to get to the same diagnostic/therapeutic destination, and as such do not have chronologically consistent viewing patterns. It is also plausible that much of what makes experts proficient is their ability to effectively process information cognitively, which may be challenging to discern accurately with gaze-tracking.

Our study has certain limitations. As a correlational study with multiple comparisons, results should be interpreted as hypothesis-generating. As a result, we limit our interpretation of this data to the general observation that higher performing residents are better able to appropriately deprioritize certain task-irrelevant information in simulation-based examinations. We propose that future studies design scenarios to incorporate *(a priori)* realistic task-irrelevant stimuli to more accurately simulate the clinical environment and better elucidate discriminating behaviours that might be indicative of expertise. These studies should utilize an experimental design to allow for causal conclusions about gaze-tracking and performance to be made. Further, we observed poor ICC/inter-rater reliability between the two individuals tabulating the results and only included one external blinded assessor for performance scores for S3 and S4. Future studies that use automated computer software (as it becomes readily available) to define exactly what constitutes a visual fixation should improve accuracy of results and speed of data generation. Finally, the authors had to make assumptions about what participants *perceived* based on what they *looked at*, without truly knowing whether their attentional resources had attended to these stimuli. Future studies could employ a mixed methods approach with post-scenario interviews to enrich the interpretation of the data (Gegenfurtner et al., 2017). Finally, our study was conducted in the simulation laboratory and, as such, we cannot make conclusions about information gathering patterns in clinical situations.

Conclusions

The results of this study suggest that there are certain visual behaviours in resuscitation-based simulations that are predictive of performance. These visual behaviours vary between cases because certain visual stimuli may be relevant in one patient presentation, but irrelevant in another. Individuals who perform better appear to have an improved ability to appropriately deprioritize task-irrelevant information and selectively process task-relevant

information. Gaze-tracking may give educators new insights into the visual processing patterns of medical learners in simulated environments.

Financial support

The authors would like to acknowledge the Kingston Resuscitation Institute for providing funding for the research assistants and access to the eye-tracking device used in this study.

Acknowledgements

The authors thank Dr. Melanie Walker for her help with proofreading this manuscript.

Appendix 1

Example QSAT

Queen's Simulation Assessment Tool (QSAT)

Station – BRADYCARDIA with PACING

Examinee Identification:_____ Date of Assessment:_____

Assessed by:_____

Primary Assessment

VITAL signs (HR/BP/O2sat/RR/Temp)
Cardiac MONITORS
IV access
Airway assessment
LOC assessment (verbal/pain/eyes), Pupils
Glucometer / Temp

1	2	3	4	5
INFERIOR Delayed or incomplete performance of all criteria	***NOVICE*** Delayed or incomplete performance of many criteria	***COMPETENT*** Delayed or incomplete performance of some criteria	***ADVANCED*** Competent performance of most criteria	***SUPERIOR*** Efficient and rapid performance of all criteria

Diagnostic Workup

History (from nurses), PMHX, Meds, Allergies
Physical Exam
ECG (interpretation)
BLOODWORK cardiac & extended lytes
BROAD DDx considered

1	2	3	4	5
INFERIOR Delayed or incomplete performance of all criteria	***NOVICE*** Delayed or incomplete performance of many criteria	***COMPETENT*** Delayed or incomplete performance of some criteria	***ADVANCED*** Competent performance of most criteria	***SUPERIOR*** Efficient and rapid performance of all criteria

Therapeutic Actions

TCP capture & recognition loss of capture
IV 1-2L NS bolus & BP support
02 by mask / TVP request
Consideration of sedation / Airway
BBLOCKER TOXICITY management (see refs)

1	2	3	4	5
INFERIOR Delayed or incomplete Performance of all criteria	***NOVICE*** Delayed or incomplete performance of many criteria	***COMPETENT*** Delayed or incomplete performance of some criteria	***ADVANCED*** Competent performance of most criteria	***SUPERIOR*** Efficient and rapid performance of all criteria

Communication

Clear and concise orders and direction
Prioritizes tasks and anticipates further steps
Demonstrates leadership in managing crisis
ICU (or CARDIOLOGY) + TOX consultation
Requests family presence

1	2	3	4	5
INFERIOR Delayed or incomplete performance of all criteria	***NOVICE*** Delayed or incomplete performance of many criteria	***COMPETENT*** Delayed or incomplete performance of some criteria	***ADVANCED*** Competent performance of most criteria	***SUPERIOR*** Efficient and rapid performance of all criteria

OVERALL PERFORMANCE

1	2	3	4	5
INFERIOR All skills require significant improvement	***NOVICE*** Most skills require moderate or significant improvement	***COMPETENT*** Some skills require moderate improvement	***ADVANCED*** Some skills require minor improvement	***SUPERIOR*** Few, if any skills require only minor improvement

Additional Comments:

Appendix 2

Average fixation data in the first 60 seconds of each scenario. Values listed are in seconds.

		Time to first vitals fixation	Total time in silence	Total vitals fixation time	Total RN fixation time	Total labs fixation time	Total ECG fixation time	Number of patient fixations	Number of RN fixations	Number of RT fixations	Number of fixations on vital signs	Number of medication list fixations
S1	Min	2.0	23.5	2.5	2.8	4.2	0.0	6.0	0.0	0.0	3.0	NA
	Max	27.0	41.0	31.0	20.3	27.1	47.3	13.0	12.5	6.5	10.5	NA
	Mean	8.1	31.1	12.2	9.8	14.5	11.8	8.9	6.5	2.1	6.0	NA
	SD	7.0	6.9	6.9	5.5	7.0	15.5	2.1	2.9	2.1	2.3	NA
S2	Min	1.8	22.0	4.0	1.2	2.0	4.0	3.0	0.0	2.0	0.0	0.0
	Max	20.5	43.0	18.6	16.0	15.8	19.0	10.5	9.0	11.0	8.0	7.5
	Mean	7.9	33.2	10.4	8.1	9.5	11.7	7.2	4.3	5.2	4.2	4.1
	SD	6.1	6.8	4.9	5.4	4.1	4.1	2.2	2.8	2.5	2.6	1.9
S3	Min	4.0	25.8	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0
	Max	233.0	50.5	20.5	18.1	41.8	13.7	13.0	9.0	6.5	9.5	9.0
	Mean	51.2	36.4	7.9	6.6	16.2	4.0	8.1	2.8	2.2	2.0	3.3
	SD	40.6	6.4	4.9	5.3	9.9	4.0	2.7	3.0	1.8	2.6	2.5
S4	Min	2.0	23.7	0.8	0.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0
	Max	131.5	54.5	21.4	25.5	53.4	18.0	18.5	12.0	9.5	8.5	9.5
	Mean	18.1	41.0	7.6	7.7	30.3	6.8	9.1	3.2	2.8	2.0	3.9
	SD	23.7	5.9	4.9	5.5	12.8	4.8	3.2	3.2	2.7	2.8	3.1

Appendix 3

Mean score by residency level and scenario

Year	Scenario 1			Scenario 2			Scenario 3			Scenario 3		
	Num.	Mean Score	SD	Num.	Mean Score	SD	Num.	Mean Score	SD	Num.	Mean Score	SD
PGY1	-	-	-	-	-	-	5	58%	.74	5	58%	.22
PGY2	1	*	-	2	51%	.11	7	68%	.16	7	60%	.18
PGY3	2	75%	.20	2	66%	.33	2	84%	.06	2	71%	.12
PGY4	1	*	-	1	*	-	3	83%	.15	3	62%	.19
Fellow	3	85%	.98	3	93%	.51	1	*	-	1	*	-
CCFP +1	4	57%	.13	4	56%	.18	10	66%	.12	10	57%	.20

*Scores removed due to concerns about anonymity of participants.

References

Bick, J. S., DeMaria Jr, S., Kennedy, J. D., Schwartz, A. D., Weiner, M. M., Levine, A. I., . . . Wagner, C. E. (2013). Comparison of expert and novice performance of a simulated transesophageal echocardiography examination. *Simulation in Healthcare, 8*(5), 329-334.

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine, 88*(6), 872-883.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363.

Gaba, D. M., Howard, S. K., Fish, K. J., Smith, B. E., & Sowb, Y. A. (2001). Simulation-based training in anesthesia crisis resource management (ACRM): a decade of experience. *Simulation & Gaming, 32*(2), 175-193.

Gegenfurtner, A., Kok, E., Geel, K., Bruin, A., Jarodzka, H., Szulewski, A., & Merriënboer, J. J. (2017). The challenges of studying visual expertise in medical image diagnosis. *Medical Education, 51*(1), 97-104.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552.

Gegenfurtner, A., Quesada-Pallarès, C., & Knogler, M. (2014). Digital simulation-based training: A meta-analysis. *British Journal of Educational Technology, 45*(6), 1097-1114.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30*(3), 304-337.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 172.

Hall, A. K., Dagnone, J. D., Lacroix, L., Pickett, W., & Klinger, D. A. (2015). Queen's Simulation Assessment Tool: development and validation of an assessment tool for resuscitation Objective Structured Clinical Examination Stations in emergency medicine. *Simulation in Healthcare, 10*(2), 98-105.

Helle, L., Nivala, M., Kronqvist, P., Gegenfurtner, A., Björk, P., & Säljö, R. (2011). Traditional microscopy instruction versus process-oriented virtual microscopy instruction: a naturalistic experiment with control group. *Diagnostic Pathology, 6*(1), S8.

Henneman, E. A., Marquard, J. L., Fisher, D. L., & Gawlinski, A. (2017). Eye Tracking: A Novel Approach for Evaluating and Improving the Safety of Healthcare Processes in the Simulated Setting. *Simuation in Healthcare, 12*(1), 51-56. doi:10.1097/sih.0000000000000192

Issenberg, B. S., Mcgaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher, 27*(1), 10-28.

Kim, J., Neilipovitz, D., Cardinal, P., Chiu, M., & Clinch, J. (2006). A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical Care Medicine, 34*(8), 2167-2174.

Kok, E. M., de Bruin, A. B. H., Robben, S. G. F., & van Merriënboer, J. J. G. (2012). Looking in the Same Manner but Seeing it Differently: Bottom-up and Expertise Effects in Radiology. *Applied Cognitive Psychology, 26*(6), 854-862. doi:10.1002/acp.2886

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1999). Expertise and tacit knowledge in medicine. *Tacit knowledge in professional practice: Researcher and practitioner perspectives*, 75-99.

Seppänen, M., & Gegenfurtner, A. (2012). Seeing through a teacher's eyes improves students' imaging interpretation. *Medical Education, 46*(11), 1113-1114. doi:10.1111/medu.12041

Szulewski, A., Brindley, P., & Van Merrienboer, J. J. G. (2017). Decision-making during medical crises. In P. Brindley & P. Cardinal (Eds.), *Crisis Resource Management in Acute Care Medicine* (1st ed., pp. 36-43). Ottawa: Royal College of Physicians and Surgeons Canada.

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L., & van Merriënboer, J. J. (2016). Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 1-18.

Szulewski, A., & Howes, D. (2014). Combining First-Person Video and Gaze-Tracking in Medical Simulation: A Technical Feasibility Study. *The Scientific World Journal, 2014*.

Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987.

Watkins, S. C., Roberts, D. A., Boulet, J. R., McEvoy, M. D., & Weinger, M. B. (2017). Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. *Simulation in Healthcare*, *12*(2), 69-75.

Chapter 5

Getting Inside the Expert's Head: An Analysis of Physician Cognitive Processes During Trauma Resuscitations

Published as: White, M. R., Braund, H., Howes, D., Egan, R., Gegenfurtner, A., van Merriënboer, J. J. G., & Szulewski, A. (2018). Getting inside the expert's head: an analysis of physician cognitive processes during trauma resuscitations. *Annals of Emergency Medicine*, *72*(3), 289-298.

Abstract

Objective
Crisis resource management skills are integral to leading the resuscitation of a critically ill patient. Despite their importance, crisis resource management (and its associated cognitive processes) have traditionally been difficult to study in the real world. The objective of this study was to derive key cognitive processes underpinning expert performance in resuscitation medicine using a new eye-tracking based video capture method during clinical cases.

Methods
Over an 18-month period, a sample of 10 trauma resuscitations led by four expert trauma team leaders were analyzed. The physician team leaders were outfitted with mobile eye-tracking glasses for each case. After each resuscitation the participant was debriefed using a modified cognitive task analysis, based on a cued-recall protocol, augmented by viewing their own first-person eye-tracking video from the clinical encounter.

Results
Eye-tracking technology was successfully applied as a tool to aid in the qualitative analysis of expert performance in a clinical setting. All participants stated that using these methods helped uncover previously unconscious aspects of their cognition. Overall, five major themes were derived from the interviews: (1) logistical awareness; (2) managing uncertainty; (3) visual fixation behaviours; (4) selective attendance to information and (5) anticipatory behaviours.

Conclusion
The novel approach of CTA augmented by eye-tracking allowed for the derivation of five unique cognitive processes underpinning expert performance in leading a resuscitation. An understanding of these cognitive processes may have the potential to enhance educational methods and to create new assessment modalities of these previously tacit aspects of expertise in this field.

Introduction

Resuscitation is a dynamic, complex, and time-sensitive field of medicine. Physicians who practice resuscitation medicine are tasked with the management of critically-ill patients, as well as leading the large multidisciplinary teams that care for them (Szulewski, Brindley, & Van Merrienboer, 2017). These facets combine to provide a uniquely complex environment in which physician leaders must operate (St.Pierre, Hofinger, Buerschaper, & Simon, 2011). The ability of a physician to expertly manage a resuscitation is dependent on a variety of parallel cognitive processes including the application of knowledge, information gathering, and decision-making (Ericsson, 2015). Identifying and categorizing these processes has been difficult and consequently our understanding of expertise, particularly within the field of resuscitation medicine, remains limited (Ericsson, 2015; Norman, Eva, Brooks, & Hamstra, 2006). Beyond the acquisition of factual or procedural knowledge, traditional notions of expert performance in resuscitation have focused on the applied principles from crisis resource management and on the avoidance of specific cognitive biases that can impact physician decision making (Carne, Kennedy, & Gray, 2012; Croskerry, 2013).

Despite the limited research focusing on expertise within resuscitation medicine, there is an extensive base of literature in a variety of other fields exploring the nature of expertise and the science of decision making (Ericsson, 2016; Kahneman, 2011; Kahneman & Klein, 2009; G. Klein, 2008; G. A. Klein, 1999). The methods used to study the detailed thought processes and decision-making techniques of experts in various fields have become known as cognitive task analysis (CTA) whereby cognition is studied in real-world contexts and professional practice at work (Crandall, Klein, Klein, & Hoffman, 2006). More recently, the methods of CTA have been applied to understand a broader range of phenomena underpinning expert performance beyond decision-making (G. Klein, 2008). The term macrocognition has been used to describe these phenomena and refers to the collection of cognitive processes and functions that characterize how people think in real-world settings. Important macrocognitive processes, derived in studying experts in other fields, include sense-making, planning, managing uncertainty, adaptation and coordination (Crandall et al., 2006; Schubert, Denmark, Crandall, Grome, & Pappas, 2013). Furthermore, the tools of CTA have also uncovered the enhanced metacognition of experts when compared to more novice performers. Metacognition has been broadly defined as “thinking about thinking” and

includes recognition of one's own memory limitations, situational awareness, and self-assessment (Crandall et al., 2006; Flavell, 1979; Ross, Shafer, & Klein, 2006).

Although the methods of CTA have been successfully utilized to uncover expert cognitive functions in domains such as firefighting and the military (Crandall et al., 2006), there is a paucity of this type of work in the field of resuscitation medicine. The high number of decisions that resuscitation physicians must make, with limited information in challenging environments, makes this field ripe for CTA and furthering our understanding of their cognitive processes.

Importantly, it has been observed that experts have limited insight into their specific cognitive processes when asked about them retrospectively, and often show both skewed and incomplete recall of events (Crandall et al., 2006; Ericsson, 2006). Several studies have attempted to overcome this difficulty by the use of head-mounted cameras as a way to generate first-person video data to supplement qualitative analysis of physician cognition (Gegenfurtner, Lehtinen, Jarodzka, & Säljö, 2017; Pelaccia et al., 2014, 2016; van Gog, Paas, van Merriënboer, & Witte, 2005).

Eye-tracking technology builds on first-person video and has been used in a variety of both non-medical and medical fields. Until recently, the application of eye-tracking technology was constrained by the bulk of the equipment, and cumbersome calibration procedures. With the advent of mobile, light, and unobtrusive glasses capable of providing detailed eye-tracking information in real-time, potential applications have blossomed. In particular, eye-tracking methodology has found a niche within medical education in studying visual expertise (Gegenfurtner, Kok, et al., 2017; Gegenfurtner & van Merriënboer, 2017). It has been used to explore visual and cognitive behaviours in surgery and radiology (Gegenfurtner, Lehtinen, et al., 2017; Kok & Jarodzka, 2017). Eye-tracking has also been used to objectively measure cognitive load (Szulewski, Gegenfurtner, Howes, Sivilotti, & van Merriënboer, 2016; Szulewski, Roth, & Howes, 2015), and to study visual fixation patterns of experts in simulation settings (Szulewski & Howes, 2014).

In the current study we present, to our knowledge, the first use of CTA augmented by participant review of video generated by eye-tracking technology during trauma resuscitations. The primary objective of this study was to gain a better understanding of the

specific cognitive processes of expert physicians while leading real-world resuscitations. A secondary objective was to demonstrate the feasibility and utility of using eye-tracking as a means of enhancing traditional CTA techniques.

Methods

Context

A phenomenological approach was used to understand the participants' experiences and awareness of their behaviours and decision making with a focus on gaining insight into the essence of medical expertise (Svensson, 1997). Phenomenological interviewing, and the techniques of CTA, focused on common elements of the lived experience of expertise of the team leader in the trauma bay while caring for an injured patient.

Participants

Four physicians were identified as local experts in resuscitation medicine and were specifically recruited, in person, to participate in this study. No potential participant approached refused nor dropped out of the study at any point. Each expert was an attending emergency medicine or critical care physician with specific fellowship training and experience in resuscitation medicine. Beyond their daily clinical practice, they all worked specifically as Trauma Team Leaders (TTLs) with a collective experience as attending trauma physicians of over 30 years. Table 1 presents information on the specific qualifications of the respective participants. Beyond these qualifications, they were also chosen based on collective consensus by the study authors as being physicians who are generally regarded, by reputation, as being excellent in their role as a trauma physician. Unfortunately, we were unable to enroll female experts as there were no women at our institution with additional fellowship-level resuscitation training acting as TTLs. This study was undertaken in the Emergency Department of a Regional Trauma Center with a large catchment area and an annual patient volume of over 60,000. Ethics approval for this study was obtained from the Queen's University Faculty of Heath Sciences Ethics Review Board (File # 6013544).

Table 1: Training/qualifications of the expert participants

Participant	Qualifications
1	Emergency medicine specialist, Trauma Fellowship
2	Emergency medicine specialist, Critical Care Fellowship
3	Emergency medicine specialist, Resuscitation Fellowship
4	Emergency medicine specialist, Critical Care Fellowship

Case Selection

Cases that activate the trauma team, by their nature, involve high patient acuity, integration of a large number of personnel, and pre-hospital notification by the local EMS system, thereby affording time in which to outfit the expert participant with the eye-tracking glasses prior to the arrival of the patient.

In total, 10 trauma cases were collected over a 15-month period between July 2015-October 2016. Two participants completed two cases each, and the other two participants each completed three cases. The participants did not control or select which cases were studied. All cases that were captured with the eye-tracking glasses were analyzed and included in this study. While completing the analysis, after seven cases, thematic saturation occurred with all themes being derived at this point in the analysis. Three further cases were analyzed, confirming data saturation, and thus case recruitment was stopped after a total of ten cases (Francis et al., 2010).

Eye-Tracking Glasses

Each participant was outfitted with Tobii Pro Glasses 2 (Figure 1) and the device was calibrated as per manufacturer recommendations. Eye movements were sampled monocularly at a rate of 50 Hz. The recording was started prior to patient arrival and continued until a natural end-point in the resuscitation at which point the participant was able to remove the eye-tracking glasses without interfering with their role as physician leader. The recording produced a first-person video with an overlying gaze indicator showing where the participant was looking in real-time (Figure 2). Each recording was uploaded to a secured study computer with Tobii Glasses Analyzer software for subsequent replay and analysis.

Figure 1: Tobii Eye-Tracking Glasses Pro 2

Figure 2: Snapshot of eye-tracking recording showing first-person video with the dynamic overlying gaze indicator.

Cognitive Task Analysis

Within 1 week of the clinical case, each participant was debriefed with the eye-tracking video recording. A CTA was used, in keeping with a phenomenological approach to qualitative analysis. Each CTA consisted of a cued-recall, retrospective protocol while the participant watched his own corresponding recording. The participants were encouraged to verbalize their internal dialogue and thought processes as they watched the replayed eye-tracking video. This method of CTA was chosen as it was thought that the use of eye-tracking glasses would trigger the experts to remember specific thought processes they had during the case. Additionally, each debrief was supplemented at the end with a pre-specified set of questions (see Appendix 1). The aim of this supplement was to collect

consistent information across the different participants, not already captured during the cued-recall portion. Finally, all participants were asked to comment on their thoughts pertaining to the utility of the eye-tracking glasses in uncovering their tacit knowledge and to describe whether the eye-tracking glasses impacted their behaviours during the resuscitation. The debriefs were between 30-45 minutes in length, the majority consisting of the cued-recall portion.

Eye-Tracking Video Analysis

The interdisciplinary research team watched the first three trauma videos together focusing on (but not being limited to) potential macrocognitive processes that we had observed in our previous resuscitation-based simulation work (Szulewski & Howes, 2014). While viewing the videos, the research team focused on observing the participants in their natural resuscitation environment. Members of the research team recorded their own individual notes for each resuscitation and then calibrated their notes following the videos. All 10 videos were then further reviewed by both the senior resident and educational doctoral student authors. The calibrated notes generated a list of specific behaviors of the expert physicians and provided additional qualitative data to the debriefs. They also provided context for the doctoral student who completed the thematic analysis of the debriefs as described below.

Thematic Analysis of Debriefs

All 10 debriefs were transcribed verbatim by two research assistants. Following data transcription, the transcripts were reviewed to correct transcription errors that were mostly related to medical jargon. The debriefs were then uploaded into ATLAS ti 1.0.5.0 software for analysis. Each debrief was analyzed separately and then analyzed across cases using an emergent thematic approach. The smallest unit of analysis (codes) were grouped into categories (consisting of multiple codes per category), which were then organized into themes (patterns) (Boyatzis, 1998; Braun & Clarke, 2006; Creswell & Creswell, 2017). These themes were observed across debriefs and experts. The videos also provided context, and a

preliminary platform to aid in the analysis of the transcripts. The research team met frequently to discuss results and emergent themes. The analysis process was iterative as codes and themes were consistently revised to best represent the data.

Research Team & Reflexivity

Data collection of the eye-tracking videos and all debriefs were completed by author MW, a senior resident completing training in both Emergency Medicine and Critical Care who had trained under each expert participant in the study. The majority of the debriefs consisted of the participants' own unprompted recollections of their thought processes while watching the eye-tracking video. During all analyses, a research logbook was used to note memos, biases, and maintain an awareness of the interpretations made. This reflective process helped the researchers to be mindful of their preconceptions about the data and helped to produce new insights stemming directly from the data (Johns, 2017). None of the derived themes were shared with participants during ongoing debriefs, nor were they involved in that process. Qualitative data analysis was conducted by an interdisciplinary research team, consisting of an attending physician, senior resident, an educational psychologist and researcher, and an education doctoral student with a background in cognition. The education research members had no medical training.

Results

Expert Cognition

Overall, five emergent themes, pertaining to the macrocognitive aspects of expert cognition, were found. These themes were: (1) logistical awareness, (2) handling uncertainty, (3) visual gaze behaviours, (4) selective attendance to information, and (5) anticipatory behaviours. Figure 3 summarizes these overall themes as well as the specific subthemes/macrocognitive processes within each. Additionally, Table 2 summarizes the themes and respective preliminary codes, and provides further supplemental quotations from the debrief transcripts.

Expert Cognition

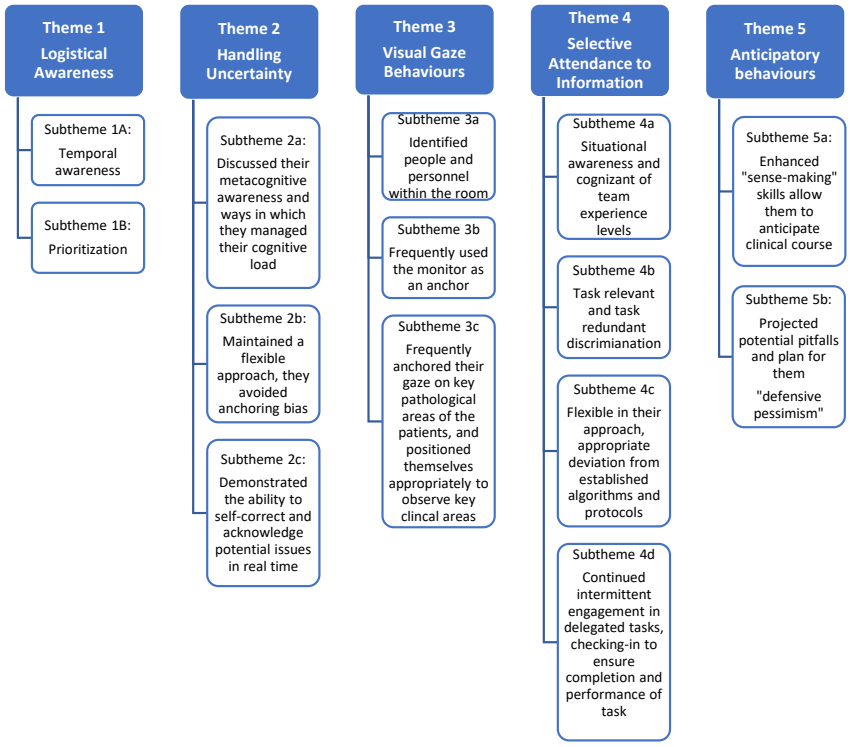


Figure 3: Summary of derived macrocognitive themes and subthemes

Table 2: Summary of themes derived from coding of primary eye-tracking videos and debrief transcriptions with supporting quotations

Themes and Subthemes	Example Codes	Example Quotations/Observations
Theme 1: logistic awareness 1a: Temporal awareness 1b: Prioritization	Attending physician role, logistic knowledge, anticipating delays, Prioritization, CT	"But as the attending, if you trust a resident—and most of our residents you do trust—to run these things, there's a lot of things you can do behind the scenes to expedite things, right?" (P3) "But at this point, I'm mainly thinking about how to prioritize getting her to the CT scanner. That's the main interest at this point." (P4)
Theme 2: handling uncertainty 2a: Metacognitive awareness, techniques for reducing cognitive load 2b: Flexible approach, avoided anchoring bias 2c: Self-correcting behaviors	Managing cognitive load, aware of need to unload, Patient context, reclassification, need for flexibility, Feeling, mad, should have called earlier	"Because I had realized, you know, the physician is doing a great job, he's explaining everything, but I have—I have, like, a number of things that I have to get done that I need to offload onto the appropriate people so things can get done." (P3) "And so what has been a casual approach to this point, I'm reclassifying in my mind, 'This as a not-so-sick trauma [patient]' to a 'This is a potentially sick trauma [patient] who may need something done pretty soon.'" (P2) "Because I knew that we were going to need ENT [ear, nose, and throat], and I was mad at myself for not having done this before." (P3)
Theme 3: Visual gaze behaviors 3a: Identified team members/personnel 3b: Used monitors to maintain situational awareness 3c: Frequent fixations on clinically relevant areas	Attention, team members, where, who, Orientation, positioning, vital signs, Attention, patient, need to check for airway injuries, under cervical collar	"And who do we have? Where are the people?" (P2) "I'm still trying to keep an eye on the vital signs. And I can actually see from here what they were." (P2) "Checking under the [cervical] spine, under the [cervical] collar for any, you know, potential airway injuries." (P4)
Theme 4: selective attendance to information 4a: Situational awareness and cognizant of team experience levels 4b: Task relevant and task redundant discrimination 4c: Appropriate deviation from protocols 4d: Checking in, frequent reengagement with delegated tasks	Awareness, team members, level of experience, Information reduction, irrelevant conversations, Attention, patient, gauging injuries, Attention, team members, checking work	"I may have delegated calling the vascular surgeon to the resident on service because partly that resident was someone I didn't feel terribly confident in, in the first case." (P1) "Um, and then when I was trying to listen to anesthesia, and then I was trying to give the story, it started to get really loud in the room, with multiple peripheral conversations, which I, you know, I deemed irrelevant at that point in time." (P3) "…[w]hereas if I determine that we don't have time, I completely deviate from the algorithm." "Really, the—one of the things I noticed is that the majority of the time, really what, uh, what I'm doing is just watching team members do their job and checking to make sure that it's going okay." (P4)
Theme 5: Anticipatory behaviors 5a: Enhanced "sense-making" skills allow them to anticipate clinical course 5b: Project potential pitfalls and plan for them ("defensive pessimism") 5c: Anticipated the need for experienced personnel	Role, predict, prepare resources, use as needed, Patient, initial priorities, pain management, difficult, Orientation, checking team members	"So I'm trying to really try to predict what's going to happen, have it ready so that if and when we need it, we utilize it." (P3) "But, uh, we're going to have a lot of problems managing, if he's not intubated, but anticipating, kind of walking the balance." (P1) "So I'm noticing here, there was an inexperienced general surgery resident from pathology." (P1)

Theme 1: Logistical awareness

Expert physicians were found to demonstrate strong awareness of the logistical aspects of managing a trauma case. Repeatedly, experts commented on the need to prioritize multiple actions, and the need to ensure both diagnostic and treatment elements were being completed as quickly as possible. Expert trauma management extends beyond just knowing what to do, but rather both prioritization and temporal awareness are necessary to enact a specific plan and see it through in the most efficient way possible. For example, Participant 3 stated:

"Classically there are delays and this patient then ends up in the emerg for x hours, their OR [operating room] gets delayed, everything gets delayed. I find that's the biggest thing that I'm doing in this case. And to do that I have to really understand the system and who to talk to." (P3)

and Participant 4 similarly stated:

"…I'm mainly thinking about how to prioritize getting her to the CT scanner." (P4)

This concept of "understanding the system" and nuances of the environment in which the expert works may be a critical component of expertise. In this way, an expert's performance may be contingent on his/her specific practice environment and may or may not be transferable to different centers with unique local practices and resources (Gegenfurtner & Szulewski, 2016). Furthermore, during the debriefs, experts frequently were aware of time, and realized the need to expedite tasks. This temporal awareness was highlighted repeatedly during each debrief and was found to be a frequent focus of the experts' cognition. Participant 1 stated:

"…you have to keep the momentum going, things happen rapidly at the outset and then people tend to lose track of time." (P1)

Furthermore, Participant 3 stated:

"…for things to actually be prepared and ready to go, so we're not waiting around for people." (P3)

Theme 2: Handling uncertainty

Participants were simultaneously able to acknowledge uncertainty, yet still make decisions, and maintain an open-framework avoiding overreliance on initial impressions (i.e. anchoring bias). Participant 3 stated:

"And then it was back to the knees and it led me to kind of question a bit whether we were doing the right thing by taking him for an angiogram, just thinking pretty clearly that it's going to be behind the knees." (P3)

and similarly Participant 2 stated:

Because the patient's clinical status evolved over the course of time, I think I was frequently re-evaluating the treatment decisions I made." (P2) and stated he was "*adjusting* [his] *therapeutic plan as* [he] *went."* (P2)

This metacognitive awareness of the participants' own uncertainty also manifested in how participants recognized and managed their cognitive load. Several participants specifically mentioned awareness of cognitive load and through the debrief realized they had deployed methods to mitigate it. Participant 2 found that "*speaking about it"* to the room helped and likewise Participant 3 stated *"if I haven't said it, I'll forget it."* Participants were also aware of the team's collective cognitive load:

"Yeah I like to kind of do short lists, I'll pick like 3 items usually, because, you know, people can only process [so much]." (P1)

Theme 3: Visual gaze behaviours

We found that experts exhibited common, specific gaze behaviours while managing trauma resuscitations. During the few minutes prior to a patient's arrival, participants were found to focus on the available personnel and equipment. For example, Participant 2 stated they were *"coming back, trying to take stock of the room again"* and in reference to the environment of the trauma bay stated *"…is there something else that's going to be an obstacle."* Their positioning in the room with respect to other team members and physical obstacles would also allow their gaze to focus on key clinical areas while assessing the patient. Participant 4 noticed positioning himself to examine a patient's airway:

"Checking under the C-spine, under the collar for any potential airway injuries." (P4)

Importantly, they all used fixation on the monitor and vital signs as a way of maintaining situational awareness if temporarily needing to fixate on a specific task or aspect of the resuscitation. Participants 4 and 2 stated respectively:

"*I go back to the monitor quite frequently, often it's something I'm just doing while I'm thinking about something else."* (P4) and

"Again I'm trying to keep an eye on the vital signs." (P2)

Theme 4: Selective attendance to information

The fourth general macrocognitive theme pertained to the participants' selective application of attentional and cognitive resources. The amount of information, visual stimuli, and decision-making density of a trauma resuscitation all serve to stress the cognitive demands of the physician leader. Participants were found to judiciously and variably apply their cognitive resources depending on the nature of the trauma and make-up of the trauma team. For example, participants delegated multiple tasks during each trauma, but in doing so, were aware of the ability of the performing personnel. Participant 1 stated:

"I may have delegated calling the vascular surgeon to the resident on service, because partly that resident was someone I didn't feel terribly confident in, in the first place." (P1)

and likewise Participant 4 stated:

"One of the things about these July resuscitations is so many of the team members are so junior and relatively unknown…well you're pretty sure they haven't got much experience and don't know what they're doing so I think I'm more attentive to what they're doing because I don't totally trust them." (P4)

This process of 'checking-in' to ensure delegated tasks were completed effectively was common amongst the participants. Importantly, the degree in which the participant monitored a delegated task was found to be contingent on the trust the leader had on the person whom they were delegating. If the participant had a high level of trust in the person completing the task, they would cognitively offload this to a greater extent than if the task had been delegated to a less competent or experienced person. For example, in one of the cases, the physician delegated the placement of a chest tube to a senior general surgical resident. Because the physician leader trusted that this was a straightforward task for the resident, few further cognitive resources were applied until after the procedure was completed, at which time the physician leader checked-in to ensure the task was completed correctly. It was also noted that very few visual fixations on the chest tube insertion procedure were made during its completion by the physician leader while he was attending to other tasks.

Participants also showed the ability to selectively process information, often ignoring, or de-emphasizing less critical information. Participant 4 noticed how he would often need to selectively interpret information fed to him by the trauma team (often comprised of junior residents) in order not to get bogged down in extraneous information.

"But [the residents] are also sometimes very keen to show that they know stuff, so they'll often tell you stuff that is not so much relevant but is more to show they know their information." (P4)

Moreover, Participant 4 stated:

"The neurosurgery resident is reporting a change in the pupils, it wasn't there before. So to be honest I'm hearing that report and not believing it. Not that I'm not believing it, I don't think there's a clinically significant difference, it's more perception. (P4)

In the context of severe traumatic brain injury, a potential change in pupil size may be a critical piece of information. The patient in question, however, was deemed not to have a clinically significant difference in pupil size by the expert participant and thus this purported change was placed in the overall context of the patient (in contrast to the junior resident reporting the change).

Theme 5: Anticipatory behaviours

The final macrocognitive theme concerned the anticipatory behaviours of participants. This related to two aspects of the trauma resuscitations. First, participants displayed strong sense-making skills, and were able to quickly get a gestalt overview of the management priorities and the likely disposition of the patient. In this way, they were able to get specific therapies and diagnostic tests organized quickly. Secondly, based on their experience and knowledge, participants were able to forecast into the future, anticipating potential pitfalls and plan for them. This "defensive pessimism" meant that participants would be prepared in the event that things went wrong, or in contrast to expectations. For example, Participant 3 ensured that a physician, experienced in performing surgical airways, was consulted and available as the patient did not have a secured airway, and had features suggesting a high-risk airway. Likewise, Participant 1 ensured the staff general surgeon (in contrast to the

junior general surgery resident who was present) was immediately contacted, as the participant projected that the patient was going to need an operation. Participant 1 stated:

"…there were a lot of things, even though again, nothing is happening here right now, and from an untrained eye, everything is fine, but it's just, I'm finding all of this medicine, all anticipating, it's all what could happen, and if it happens, do we have a plan? Do we have this?" (P1)

Eye-Tracking Enhanced CTAs

The use of the eye-tracking glasses was found to be seamless in these real-world clinical cases. None of the participants removed the glasses during the cases. Participants denied any significant discomfort associated with the device, and they did not feel that it impeded patient care or altered their management of the case. For example, P2 stated:

"Around the time I realized the patient was sick, I stopped being aware that I was wearing the glasses." And P1 similarly stated:

"[The eye-tracking glasses] did not affect my decision making or performance. "

Finally, while reviewing their respective eye-tracking videos, all participants mentioned that this technique enabled to them to more accurately recall their specific thought processes during the resuscitation.

Discussion

This study represents the novel application of mobile eye-tracking technology to enhance the ability of CTA to uncover cognitive elements underpinning expert performance in resuscitation medicine. To our knowledge, this represents both the first application of this technology to CTA, and the first foray into elucidating expert cognitive performance specifically in real-world resuscitation medicine.

Figure 3 summarizes the five derived themes from the qualitative analysis including logistical awareness, managing uncertainty, visual gaze behaviours, selected cognitive attention, and anticipatory behaviours. In another study, Schubert et al. (2013)

characterized the macrocognitive differences in experts compared to novices while working in the emergency department. Similar cognitive themes pertaining to expertise were derived in that study. The present work builds on the work of Schubert et al. (2013) in several important ways. First, we utilized eye-tracking glasses, which provided not only first-person video, but additionally showed exactly where the participants were looking at all times from their own perspective. Furthermore, this work involved active resuscitations requiring the leadership of large clinical teams, which differ from the less acute cases described by Schubert. This creates important differences in how the themes are applied.

Much of the previous work highlighting the importance of cognitive processes in resuscitation medicine has focused on borrowed principles of crisis resource management. These cognitive processes and human factor principles such as situational awareness, cognitive load management, and information processing have been previously derived in studies of other fields, particularly aviation (Flin & Maran, 2004). Although some of these principles have been applied to resuscitation medicine, particularly in a simulation environment, the current study represents the first time they have been studied in real-world resuscitation cases from a first-person perspective. We believe that this study represents a novel contribution to the field, by deriving these processes from real-world data and in doing so, showing how they are directly applied in specific real-life cases, beyond theoretical and simulated environments.

This work has important applications to both education and assessment. Zupanc, Burgess-Limerick, and Wallis (2015) utilized video recording and CTA techniques to derive a competency framework for colonoscopy. Using this methodology, they were able to identify twenty-seven real-world-derived competency components, providing a principled structure for future training programs and the design of better formative assessments. Similarly, this work provides a basis for which to enhance pre-existing educational and assessment techniques. Within the field of resuscitation medicine, simulation has been relied on heavily to train learners in tacit areas of performance (Petrosoniak & Hicks, 2013; Szulewski et al., 2018). The authors suggest that based on the results of this study, some of these previously tacit areas have now been better described. Given this, there is the possibility of utilizing the results from this study to inform more research that could, in turn, improve simulation education.

For example, to highlight the concept of selective attendance to information (Theme 4), simulation scenarios could purposely present extraneous information to challenge advanced learners to decide on their relevance in the management of a scenario. Selective cognitive attention demonstrates the ability of experts to place information in its appropriate context. This aligns well with the information reduction hypothesis (Haider & Frensch, 1996, 1999) which holds that experts in a given field are able to distinguish between task-relevant and task-redundant information. Thus, the improved performance of experts may at least be partially attributable to their ability to decrease the amount of irrelevant information that they have to process. Based on the results of this study, our group has begun incorporating these types of "distractors" in simulation education and has found that better performing residents are better able to appropriately discard task-irrelevant clinical stimuli while focusing on more relevant stimuli which may have an impact on patient management.

A particularly germane application of this work is the way in which trainees are assessed. Currently, medical education is transitioning to a competency-based framework of assessment. Competency-based medical education focuses on specific predefined abilities and activities in which a trainee should be able to perform (Frank et al., 2010). Traditional formal assessment paradigms emphasize assessment of mostly factual knowledge. Similar to Zupanc et al. (2015), by using real-life resuscitations to derive a set of cognitive processes characterizing expert performance, these processes could be applied in designing more rigorous assessment tools.

In addition to our primary objective of deriving cognitive processes of experts, this study also demonstrates the utility of cued-recall augmented by eye-tracking as a method of enhancing real-world CTA. The use of eye-tracking allowed the participants to more easily recall details of their internal thought processes during each trauma case. The first-person video generated with the gaze indicator allowed each participant to "go back in time" and retrace their cognitive steps. This yielded a rich amount of data to be analyzed, which would not have been previously available for study. Beyond the enhancement of CTA, this study demonstrates other potential applications for eye-tracking in real-world resuscitation medicine. Since the participants found the eye-tracking glasses to be minimally invasive and did not impede their ability to manage trauma resuscitations, we feel that the methodology described herein could be applied to other clinical cases in various contexts for research as

well as potentially for quality improvement purposes and critical incident reviews. In addition, after appropriately addressing privacy issues, first-person videos generated during clinical cases could be used during educational rounds as a means to coach junior learners as well as focal points for discussion during continuing medical education sessions.

The current work has several limitations. The relatively small sample size of 4 expert physicians and 10 trauma cases potentially limits the generalizability of our results. This being said, we reached qualitative data saturation which makes the emergence of additional themes less likely. Moreover, this research was limited to a single center and it is possible that derived themes may relate to shared traits resulting from participants' similar work environments and experiences. Another limitation is the fact that all our participants were male, and therefore we cannot rule out that female physician leaders may have slightly different approaches to crisis resource management that we did not observe (Dayal, O'Connor, Qadri, & Arora, 2017). This work also only examined experts, and did not directly compare their cognitive processes with those of novices. Future work should involve experts at different sites, team leaders of varying experience and training and clinical cases beyond traumas such as airway management and cardiac arrest.

Conclusion

The cognitive processes of expert physicians leading real-world trauma resuscitations were studied using the novel combination of eye-tracking glasses with the qualitative tool of cognitive task analysis. Overall, five cognitive themes were derived including: (1) logistical awareness (2) handling uncertainty (3) visual gaze behaviours (4) selective attendance to information and (5) anticipatory behaviours. Furthermore, we found that cued-recall augmented by eye-tracking technology was a practical, unobtrusive and useful technique for cognitive task analysis in the resuscitation bay. With more study, these tools could enrich simulation education and provide a framework to aid the assessment of previously tacit aspects of expertise.

Acknowledgments

The authors would like to thank Dr. Marco Sivilotti for his insights and help with proofreading.

Appendix 1

Questions asked during the semi-structured interview at the end of each debrief

1. How much notice were you given prior the patient arrival in the trauma bay?
2. Was the entire trauma team present?
3. Were you given sufficient information by EMS prior to patient arrival to formulate potential management priorities?
4. Was the patient stabilized at a peripheral hospital prior to transfer?
5. Did EMS manage to secure IV access/airway prior to arrival?
6. Did you have a chance to brief/meet with the trauma team prior to patient arrival?
7. When the patient arrived, what were your initial priorities?
8. At what point were you able to confidently assess the patient's stability and potential injuries?
9. What information did you base this on?
10. At what point were you able to confidently determine the management priorities for the patient?
11. What information did you base this on?
12. Were any immediate procedures required in the trauma bay (e.g. intubation, chest tube insertion)?
13. Did you feel that the trauma team worked well together?
14. To what extent did you feel that individual components of the trauma team (nursing, anesthesia, general surgery, neurosurgery, orthopedic surgery) performed at an acceptable/exceptional level?
15. Overall, how critically ill was the patient?
16. How confident were you in the management of this patient?
17. Did you feel that were any extra distractions impeding your ability to manage this trauma?
18. Do you feel you were able to maintain situational awareness during management of the trauma?
19. Overall, did you feel the trauma was run smoothly and effectively?
20. Was there anything that could have been improved in the execution of trauma resuscitation for this patient?
21. To what extent did you notice that you were wearing the Tobii © glasses during the trauma?

References

Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, CA, US: Sage Publications, Inc.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi:10.1191/1478088706qp063oa

Carne, B., Kennedy, M., & Gray, T. (2012). Review article: Crisis resource management in emergency medicine. *Emergency Medicine Australasia, 24*(1), 7-13. doi:10.1111/j.1742-6723.2011.01495.x

Crandall, B., Klein, G., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*: Mit Press.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*: Sage publications.

Croskerry, P. (2013). From Mindless to Mindful Practice — Cognitive Bias and Clinical Decision Making. *New England Journal of Medicine, 368*(26), 2445-2448. doi:10.1056/NEJMp1303712

Dayal, A., O'Connor, D. M., Qadri, U., & Arora, V. M. (2017). Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. *JAMA Internal Medicine, 177*(5), 651-657.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. *The Cambridge handbook of expertise and expert performance*, 223-241.

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: a perspective from the expert-performance approach with deliberate practice. *Academic Medicine, 90*(11), 1471-1486. doi:10.1097/acm.0000000000000939

Ericsson, K. A. (2016). *Peak: How to master almost anything*: Penguin.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906-911. doi:10.1037/0003-066X.34.10.906

Flin, R., & Maran, N. (2004). Identifying and training non-technical skills for teams in acute medicine. *BMJ Quality & Safety, 13*(suppl 1), i80-i84.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health, 25*(10), 1229-1245. doi:10.1080/08870440903194015

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., . . . Dath, D. (2010). Competency-based medical education: theory to practice. *Medical Teacher, 32*(8), 638-645.

Gegenfurtner, A., Kok, E., Geel, K., Bruin, A., Jarodzka, H., Szulewski, A., & Merriënboer, J. J. (2017). The challenges of studying visual expertise in medical image diagnosis. *Medical Education, 51*(1), 97-104.

Gegenfurtner, A., Lehtinen, E., Jarodzka, H., & Säljö, R. (2017). Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. *Computers & Education, 113*, 212-225. doi:https://doi.org/10.1016/j.compedu.2017.06.001

Gegenfurtner, A., & Szulewski, A. (2016). Visual expertise and the Quiet Eye in sports – comment on Vickers. *Current Issues in Sport Science, 1*(1).

Gegenfurtner, A., & van Merriënboer, J. J. (2017). Methodologies for studying visual expertise. *Frontline Learning Research, 5*(3), 1-13.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30*(3), 304-337.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 172.

Johns, C. (2017). *Becoming a reflective practitioner*: John Wiley & Sons.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515.

Klein, G. (2008). Naturalistic decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 456-460.

Klein, G. A. (1999). *Sources of power: How people make decisions*: MIT press.

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education, 51*(1), 114-122.

Norman, G., Eva, K., Brooks, L., & Hamstra, S. (2006). Expertise in Medicine and Surgery. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 339-354). Cambridge: Cambridge University Press.

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Dory, V., & Charlin, B. (2014). How and When Do Expert Emergency Physicians Generate and Evaluate Diagnostic Hypotheses? A Qualitative Study Using Head-Mounted Video Cued-Recall Interviews. *Annals of Emergency Medicine, 64*(6), 575-585. doi:https://doi.org/10.1016/j.annemergmed.2014.05.003

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Dory, V., & Charlin, B. (2016). From Context Comes Expertise: How Do Expert Emergency Physicians Use Their Know-Who to Make Decisions? *Annals of Emergency Medicine, 67*(6), 747-751. doi:https://doi.org/10.1016/j.annemergmed.2015.07.023

Petrosoniak, A., & Hicks, C. M. (2013). Beyond crisis resource management: new frontiers in human factors training for acute care medicine. *Current Opinion in Anesthesioly, 26*(6), 699-706. doi:10.1097/aco.0000000000000007

Ross, K. G., Shafer, J. L., & Klein, G. (2006). Professional judgments and “naturalistic decision making”. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 403-419).

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A., & Pappas, J. (2013). Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department. *Annals of Emergency Medicine, 61*(1), 96-109.

St.Pierre, M., Hofinger, G., Buerschaper, C., & Simon, R. (2011). The Challenge of Acute Healthcare. In *Crisis Management in Acute Care Settings: Human Factors, Team Psychology, and Patient Safety in a High Stakes Environment* (pp. 23-39). Berlin, Heidelberg: Springer Berlin Heidelberg.

Svensson, L. (1997). Theoretical Foundations of Phenomenography. *Higher Education Research & Development, 16*(2), 159-171. doi:10.1080/0729436970160204

Szulewski, A., Braund, H., Egan, R., Hall, A. K., Dagnone, J. D., Gegenfurtner, A., & van Merrienboer, J. J. (2018). Through the Learner's Lens: Eye-Tracking Augmented Debriefing in Medical Simulation. *Journal of Graduate Medical Education, 10*(3), 340-341.

Szulewski, A., Brindley, P., & Van Merrienboer, J. J. G. (2017). Decision-making during medical crises. In P. Brindley & P. Cardinal (Eds.), *Crisis Resource Management in Acute Care Medicine* (1st ed., pp. 36-43). Ottawa: Royal College of Physicians and Surgeons Canada.

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L. A., & van Merriënboer, J. J. G. (2016). Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 1-18. doi:10.1007/s10459-016-9725-2

Szulewski, A., & Howes, D. (2014). Combining First-Person Video and Gaze-Tracking in Medical Simulation: A Technical Feasibility Study. *The Scientific World Journal, 2014*.

Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987.

van Gog, T., Paas, F., van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the Problem-Solving Process: Cued Retrospective Reporting Versus Concurrent and Retrospective Reporting. *Journal of Experimental Psychology: Applied, 11*(4), 237-244. doi:10.1037/1076-898X.11.4.237

Zupanc, C. M., Burgess-Limerick, R., & Wallis, G. (2015). Strategy influences directional control–response compatibility: evidence from an underground coal mine shuttle car simulation. *Theoretical Issues in Ergonomics Science, 16*(1), 1-19. doi:10.1080/1463922X.2013.857738

Chapter 6

Starting to think like an expert: an analysis of resident cognitive processes during simulation-based resuscitation examinations

Published as: Szulewski, A., Braund, H., Egan, R., Gegenfurtner, A., Hall, A. K., Howes, D., Dagnone, D., van Merriënboer, J. J. G. (In Press). Starting to think like an expert: an analysis of resident cognitive processes during simulation-based resuscitation examinations*. Annals of Emergency Medicine.*

Abstract

Introduction
Simulation is commonly used to teach and assess crisis resource management (CRM) skills in emergency medicine (EM) residents. However, our understanding of the cognitive processes underlying CRM skills is limited as these processes are difficult to assess and describe. The objective of this study was to uncover and characterize the cognitive processes underlying CRM skills and to describe how these processes varied between residents based on performance in a simulation-based examination.

Methods
Twenty-two of 24 eligible EM trainees from one tertiary academic centre completed one or two resuscitation-based examinations in the simulation laboratory. Resident performance was assessed by a blinded expert using an entrustment-based scoring tool. Participants wore eye-tracking glasses that generated first person video that was used to augment subsequent interviews led by an EM faculty member. Interviews were audio recorded and then transcribed. An emergent thematic analysis was completed using a codebook that was developed by four research assistants, with subsequent analyses conducted by the lead research assistant with input from EM faculty. Themes from high and low performing residents were subsequently qualitatively compared.

Results
Higher performing residents were better able to anticipate, selectively attend to relevant information, manage cognitive demands, and took a concurrent (as opposed to linear) approach to managing the simulated patient.

Conclusion
The results provide new insights into residents' cognitive processes while managing simulated patients in an examination environment and how these processes vary with performance. More work is needed to determine how best to apply these findings to improve CRM education.

Introduction

The practice of Emergency Medicine (EM) is dependent on effective and efficient decision-making. The environment of the Emergency Department (ED), however, with its frequent interruptions, distractions and time pressures, makes decision-making challenging and error-prone (Croskerry & Sinclar, 2001; Schubert, 2013). Despite these challenges, emergency physicians are tasked with regularly making critical decisions that may have life or death consequences for patients (Helmreich, 2000; Sundar, 2007). Although decision-making itself is central to patient care, the broader scope of crisis resource management (CRM) skills, including leadership, situational awareness, communication skills, and resource utilization are integral to providing effective patient care in the ED (Hicks, 2008; Kim, 2006). This is especially true during resuscitation cases where the physician team leader simultaneously manages the patient, the available resources, and an interdisciplinary healthcare team (McIntyre, 1995; Salas, 2007; Schull, 2001).

CRM skills are often taught and practiced by learners in simulation contexts, where trainees can act as team leaders, while both honing and demonstrating their CRM skills safely (Johnson, 1994; Reznek, 2003). The simulation environment provides trainees with the opportunity to develop interprofessional team skills and to critically examine behaviours that support or detract from team performance (Cook, 2018). However, the extent to which medical simulation allows learners to develop the cognitive processes that underpin the CRM skills required of practitioners has been insufficiently investigated (Hagiwara et al., 2016). Importantly, previous studies rely on post hoc interviews based upon physicians' memories of simulated events (Mills et al., 2016), which degrade over time and may lack detail.

Although some elements of CRM performance are observable, (and therefore able to be assessed and studied with currently available tools), other elements, such as the cognitive processes that underlie CRM skills, are not (White et al., 2018). Our understanding of these *invisible* cognitive processes remains limited. One reason is that as individuals gain expertise in a domain, a greater proportion of their decision-making becomes tacit, which makes explaining their thought processes challenging (K.A. Ericsson, 2006). Further, since the clarity of *in situ* CRM memories degrades over time, individuals have difficulty remembering

and analyzing subsequent events. As such, it is difficult to determine if simulated CRM scenarios evoke cognitive approaches similar to those of medical crises in real world settings.

In studying trauma team leaders in real clinical settings, we have previously demonstrated that a post-hoc cognitive task analysis augmented by review of first person video generated by an eye-tracking device is a useful technique to uncover these cognitive processes (White et al., 2018). In that study, expert trauma team leaders were found to demonstrate logistic awareness, manage uncertainty, strategically direct their gaze, selectively attend to information, and exhibit anticipatory behaviours while treating critically injured patients. The use of eye-tracking technology was particularly important as it allowed participants to "relive" their experience from their own perspective, emphasizing their gaze fixations, thus allowing for more complete recall of events and cognitive processes (Gegenfurtner et al., 2017; Kok & Jarodzka, 2017).

To address whether simulation as a tool to develop CRM skills provides value, there is a need to describe how CRM-related cognitive processes are activated in trainees through simulation. Given the wide variability in trainee performance in simulations, an understanding of how these cognitive processes might differ based on performance is necessary. Subsequently examining the relationship between these simulation-derived cognitive processes with those derived in real clinical settings may provide additional evidence for the utility of simulation as an educational modality in teaching CRM skills to medical trainees.

In the current study, we were interested in uncovering and describing the cognitive processes of EM trainees during simulation-based resuscitation examinations using a qualitative approach and describing how these processes varied between participants stratified according to examination performance.

Methods

This study received ethical approval through the institutional health sciences research ethics board. Participants gave informed consent and did not receive compensation.

Study Design and Setting

This qualitative study used a phenomenological approach in an effort to better understand residents' experiences and cognitive processes underpinning their CRM behaviours as they acted as physician team leaders during simulation-based examinations. Phenomenological qualitative approaches to research seek to describe how human beings experience lived phenomena (Davidsen, 2013). This approach was chosen as it aligned with the subsequently described cognitive task analysis and we felt it was best suited to answer this type of research question (as opposed to more traditional positivistic research methods). We used the consolidated criteria for reporting qualitative research (COREQ) checklist (Tong, Sainsbury, & Craig, 2007) to ensure methodological rigour (See Appendix A). As outlined below, quantitative performance data was used to stratify participants for subsequent analyses. All data were collected at our Clinical Simulation Centre.

Participant Selection

All EM residents participating in their mandatory biannual simulation-based objective structured clinical examinations (OSCEs) from our Canadian academic tertiary care teaching centre were invited to participate in the study. Residents were enrolled in either the Royal College of Physicians of Canada (FRCP) program, the College of Family Physicians of Canada - Emergency Medicine (CCFP-EM) program, or the Resuscitation Fellowship program. In Canada, emergency physicians are trained in one of two streams: residents in the FRCP program complete a 5-year residency and become EM specialists, while residents in the CCFP-EM program complete an additional year of EM training after completing a 2-year family medicine residency. Trainees in the resuscitation fellowship are senior residents or attending physicians who complete an additional year of resuscitation-focused training.

Participants were recruited in-person once they arrived at their examination. All participants were familiar with the simulation lab and the examination format.

Protocol for Simulation Sessions

Trainees participated in one or two simulation-based examinations (based on scheduling and availability), each consisting of two scenarios. The first was conducted in August 2016 (COPD Exacerbation and GI bleed), and the second in February 2017 (VFib / STEMI and Hyperkalemia / Bradycardia). The format of this simulation-based resuscitation examination and an accompanying validity argument for its use as an assessment process have been described elsewhere (Hagel, Hall, & Dagnone, 2016; Hall, Dagnone, Lacroix, Pickett, & Klinger, 2015).

In each examination, trainees managed a simulated patient while leading a team consisting of a nurse actor and a respiratory therapist actor. Please refer to **Figure 1** for a visual representation of the protocol. Before entering the room, trainees read a clinical vignette outlining relevant clinical context (See Appendix B for an example of one such clinical vignette). To increase the functional task alignment of the simulated scenarios with real clinical work and to assess residents' CRM skills, realistic distractors were embedded, containing stimuli which a high-performer would generally deprioritize (A. Szulewski et al., 2019). For example, in one case, the nurse provided an irrelevant high glucose reading. In another, participants were tasked with answering a phone call from a peripheral hospital where the caller was unnecessarily longwinded.

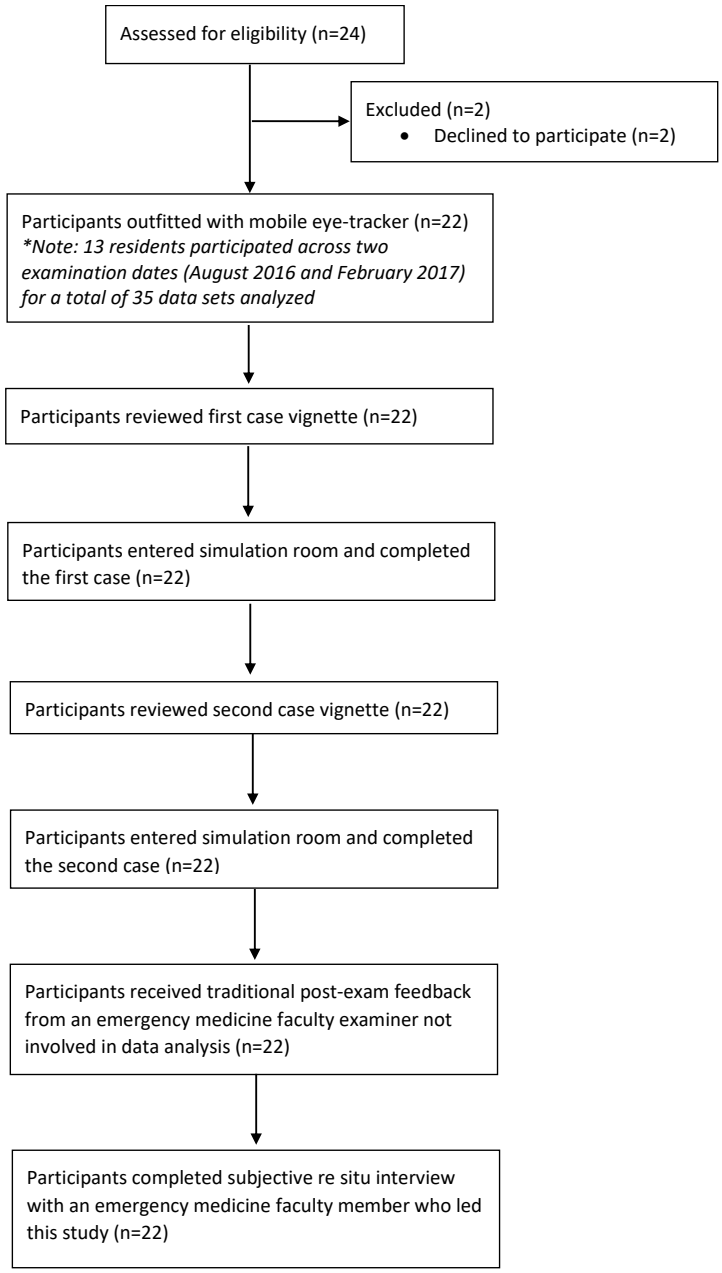


Figure 1: Overview of study protocol.

Participants wore Tobii© eye-tracking glasses (Tobii Pro Glasses 2, Danderyd, Sweden). Output from the glasses generated a first-person video with a superimposed gaze indicator that was later used during the study interview. The own-point-of-view video generated by

this tool has been shown to stimulate recall, leading to more explicit descriptions of decision-making resulting from improved psychological immersion of participants (Pelaccia, Tardif, J., Triby, & Charlin, 2017) Figure 2 shows a screenshot of video output.

Figure 2: Screenshot of video output from the eye-tracking glasses. The white circle shown is a superimposed gaze indicator that demonstrates where the participant is looking at a particular moment.

Feedback and Cognitive Task Analysis

Following case completion, trainees were given the usual post-examination feedback on their performance by an attending EM examiner (known to the participants) for 10-15 minutes. This feedback consisted of discussing the trainees' strengths as well as constructive feedback related to decision-making, teamwork and communication. After this traditional feedback session, residents completed an additional 30-minute individual debriefing session (the focus of this study). During this session, another emergency physician (author AS) guided residents through a cognitive-task analysis using a subjective re situ interview protocol (Pelaccia, Tardif, J., Triby, & Charlin, 2017), where residents viewed their first-person video performance, starting when they began to read the first clinical vignette. AS had prior experience with this type of interviewing technique. The video was played and

then paused intermittently by the interviewer at critical junctures and resident actions (at the interviewer's discretion) as well as during the introduction of each distractor. Though the timing of the video clips differed between participants based on their actions, the segments viewed were generally similar across participants. During video-replay, residents were asked to describe their thinking and to verbalize their thoughts including decision-making priorities, motivations, and choices. To encourage discussion through retrospective verbalization and probing techniques, questions like "describe your thinking" and "walk me through your thoughts here", were used. These interviews were audio-recorded and transcribed verbatim. In total, 35 interviews were generated (n.b. 13 of the 22 participants completed examinations and debriefs during both time periods). Residents' reflections on the utility of this session are described elsewhere (A. Szulewski, Braund, H., Egan, R., Hall, A., Dagnone, D., Gegenfurtner, A., Merrienboer, J. V, 2018).

Quantitative Analysis

Each resident's performance was recorded using a ceiling-mounted system (Kb Port, Allison Park, PA) and was then scored by an external blinded expert reviewer with no knowledge of the trainees, their training level or the purpose of the study. The blinded external rater used an entrustment-based (Ten Cate, 2013) scoring tool, termed the Resuscitation Assessment Tool (RAT) (Weersink, Hall, Rich, Szulewski, & Dagnone, 2018) (available in Appendix C) which was developed as a modification of the previously derived Queen's Simulation Assessment Tool (QSAT) which has validity evidence in a similar context. Strictly speaking, the RAT itself has not been directly validated but it was deemed to be more contextually appropriate given its widespread use in resuscitation assessment in the EM residency program at our institution. The expert assessor received an orientation training session which involved rating a standardized sample of training video recordings of varying levels of performance using the RAT and then reviewed with one of the investigators (AKH) until consensus scoring was achieved. Note that some of the RAT scores obtained during this study were subsequently used as part of a larger data set for an unrelated study (Weersink et al., 2018).

An average performance score was calculated for each participant using their entrustment scores from the external blinded reviewer. We considered EM residents with average scores on the RAT of higher than or equal to 3.0 (corresponding to “indirect supervision” or better on the entrustment scale) the high performers, and those receiving scores lower than 3.0 the low performers. This division was used in the subsequently described qualitative analysis and themes were compared between these two groups.

Qualitative Analysis

Data analysis was conducted by non-physician team members who were at arms-length from the participants; the researchers had no knowledge of participants’ performance scores or level of training. These researchers met regularly with the physician researchers to discuss data interpretation. All interviews were transcribed verbatim. ATLAS ti 1.0.5.0 software was used to generate the codebook, code each transcript, categorize the codes, track memos, and to generate a file report detailing all codes used with quotations. As part of the coding process, four researchers generated the initial codebook by coding four interviews independently and comparing their codes. Inter-coder agreement was found when codes were the same (95% of the coded quotations across 4 interviews). The 95% level of agreement represents complete agreement across researchers where the codes had the same meaning. For the other 5%, researchers discussed the coded quotations and made changes once agreement was reached. The consensus-built codebook was used for all remaining qualitative analyses. Following the inter-coder reliability check, one researcher analyzed the remaining qualitative debriefs through a phenomenological lens (focusing on understanding participants’ experiences) and maintained a critical stance by journaling regularly and making notes about underlying assumptions. (All of the researchers involved in the initial coding process journaled during preliminary coding but three were not involved in the remaining analyses as they were only engaged to provide a critical external perspective and check inter-coder reliability). Although data saturation occurred after 25 debriefs, all 35 debriefs were analyzed.

Interviews were coded using a thematic and emergent technique (Braun & Clarke, 2006) through a descriptive phenomenological lens (Creswell, 2003; Giorgi, 2009). Upon completion of the first three transcripts, EM team members discussed the themes and provided interpretation. The smallest unit of analysis was the code, which were grouped to form categories; multiple categories were combined forming broader themes. Language used by participants was maintained throughout coding to ensure that the interpretations were representative of participants' thoughts.

Following emergent analysis, axial coding was used to find transactional themes across transcripts (Glaser & Strauss, 1967; Strauss & Corbin, 1990). We also conducted structured coding for anticipatory behaviour. To provide temporal context, the transcripts were divided into three stages: orientation, diagnostic, and therapeutic. Although grouped separately, these stages were not expected nor found to be chronologically sequential, and the iterative nature of these activities was respected. Peer debriefing was conducted with emergency medicine researchers who had not been involved in the analyses. These sessions were used to ensure that interpretations of the data were representative and appropriate.

Research Team and Reflexivity

The lead clinician-researcher (author AS) conducted the interviews. AS is a full-time faculty member in emergency medicine, is fellowship trained in resuscitation medicine and is a PhD student in health professions education. All study participants were known to AS and had previous experience working with him in a clinical context. Author HB led the data analysis. HB is a PhD student in education with background in assessment and cognition as well as experience in qualitative emergency medicine research. She had no personal experience with any of the study participants. The remaining researchers on the team had backgrounds in emergency medicine, simulation, and education.

Results

Recruited Participants

Twenty-two of 24 eligible residents opted to participate in the study. Of the 22 residents, 10 were male, 12 were female. Thirteen were FRCPC residents, 7 were CCFP-EM residents, and 2 were resuscitation fellows. Thirteen residents participated across the two examination dates.

Quantitative Findings

In total, 11 residents were categorized as high performers (mean RAT score ≥ 3.0) and 11 residents were categorized as low performers (mean RAT score < 3.0). Resident performance (high or low) was an average across examinations. For the 11 low performers, the average entrustment score out of 5 ranged from 1.50 to 2.63. For the 11 high performers, the average entrustment score ranged from 3.00 to 4.50. Our reliability analysis indicated acceptable internal consistency with a Cronbach's alpha of .730.

Qualitative Findings

Results were organized according to the four emergent themes (**Figure 3**). **Table 1** provides a summary of the codes with examples. 'P' represents participant, 'H' indicates a high performer and 'L' indicates a low performer. For additional quotations in relation to each theme please refer to Appendix D.

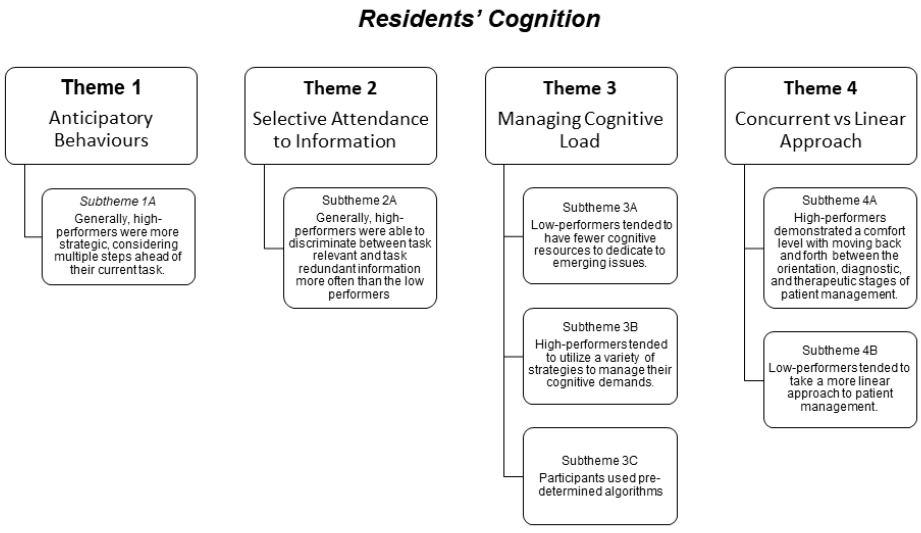


Figure 3: Summary of themes and subthemes derived

Table 1: Summary of themes, subthemes, categories, codes and example quotes

Themes and Subthemes	Example Categories	Example Codes	Low-Performers	High-Performers
Theme 1: Anticipatory Behaviours *Subtheme 1A:* Generally, high-performers were more strategic, considering multiple steps ahead of their current task.	Strategic	(Anticipatory Behaviour-Advanced) (Anticipating Need for Pacing) (Anticipatory Behaviour-Moderate) (Anticipatory Behaviour-Strong)-(Need for Intubation) (Anticipatory Behaviour-Weak)	"So, there I was just thinking to myself, you know, this is a guy who is altered and is possibly going to need intubation." (P6, ES = 2.63, Hyperkalemia / Bradycardia Case)	"… I'm thinking ahead of the game. I'm like this guy looks really sick, he's got COPD, he's old. This guy may not bounce back just with supportive care so I need to know if we're going to go all the way through to intubation or not. So, I have that kind of in the back of my head which direction I'll be taking him if he continues to get sick." (P1, ES= 3.25, COPD Case) "…[he's] losing level of consciousness, [and] then he's not protecting his airway, so I decide that it's time to actually make more full preparations. I probably should've right at the start said can we have ketamine and succinylcholine or rocuronium ready just in case we're going that way." (P3, ES = 3.00, COPD Case)

Themes and Subthemes	Example Categories	Example Codes	Low-Performers	High-Performers
Theme 2: Selective Attendance to Information *Subtheme 2A:* Generally, high-performers were able to discriminate between task relevant and task redundant information more often than the low performers.	Ignored Distractors	(Information Reduction) (Critical Venous Oxygen Saturation)-(Distractor)-(Avoided Fixation) (EMS Long History)-(Distractor)-(Cut Off) (Nephrology Phone Call)-(Distractor)-(Avoided Fixation) (Phone Call-Avoided)	"I [kind of] just tucked that away in my head if we had to intubate him, just be careful with using succinylcholine" (P14, ES = 2.50, Hyperkalemia / Bradycardia Case)	"Oh she told me the glucose, and I was like that's not important. So fleetingly looked at her… (laugh)". (P12, ES = 3.50, COPD Case) "So, [the nurse] had a phone call that she was trying to get my attention for. There was a physician from Perth that wanted to have a conversation with me. Going through my mind at the moment, it was the most critical part of the resuscitation, with me literally just about to have a look and having him fully ketaminized so I shut it down pretty quickly." (P19, ES = 4.50, COPD Case)
	Fixated on Distractors	(Scenario distractions-fixated) (Acetylcholinesterase)-(Fixated)-(Confused With Other Deficiencies) (ECG-Fixated) (Left-Bundle Block-Fixated)	"So the certain things, certain decisions that I made that I think were based on things that I got distracted by. By things that came up as the situation unfolded that maybe led me astray a little bit. And I think a lot of that is reflective to my level of training and not being comfortable to let things go as they come up and maybe over processing it. Like the ECG and the glucose at one point." (P2, ES= 1.75, COPD and GI Bleed) "I wasn't sure if I was supposed to manage this person on the phone then I quickly realized he was telling me that a somewhat stable patient needs a [computed tomography scan]. And I probably could have got there quicker and said I can't deal with this right now" (P9, ES = 2.25, COPD Case)	"Yeah. Right so this was the point where I was trying to decide that, that's tachycardic, is he going, do I think it's actually just sinus tachy. ..., it looks like he's going too fast, so then I started to get a bit confused, is this actually just a narrow complex tachycardia, is it a flutter, is it a fib." (P3, ES = 3.00, COPD Case) [regarding the acetylcholinesterase deficiency distractor] "I spent a little bit of mental energy trying to think of things that that could have meant. And it was like 'Ok can't give him [succinylcholine]. Could this in some way cause hyperkalemia?' And couldn't come up with anything." (P5, ES = 3.25, Hyperkalemia / Bradycardia Case).

Themes and Subthemes	Example Categories	Example Codes	Low-Performers	High-Performers
Theme 3: Managing Cognitive Load *Subtheme 3A*: Low-performers tended to have fewer cognitive resources to dedicate to emerging issues.	Overloaded	(Experiencing Cognitive Load) (Diagnostic Stage)-(B Lines)-(Overloaded) (Therapeutic Actions)-(Overloaded)-(Delayed Management)	"… I guess it's large cognitive load. And in you're thinking about things, I mean, your ABC is the thing, the immediate. It's very distracting and glamorous to be thinking about CPR, right, and about the resuscitation, and thinking about the underlying cause, especially when you know that the underlying cause is hypovolemia, and you've begun treating it. It can be sort of easy to focus on getting circulation back because what does it matter if there's blood coming out if the heart isn't pumping?" (P11, ES = 2.00, GI Bleed Case)	"And I think I was just thinking about that so much that I…stopped at ordering the hemoglobin, and didn't think about all the other things…I was so stuck. I definitely got cognitively stuck on the hemoglobin and whether… I was trying to decide resuscitation vs. GI bleed treatment… And so I think I felt disorganized on this one..." (P3, ES = 3.00, GI Bleed Case)
Subtheme 3B: High-performers tended to utilize a variety of strategies to manage their cognitive demands.	Cognitive Flexibility	(Managing Cognitive Load) (Cognitive Reminder)		"That proved a bit difficult in my mind, though. Kind of figured out, I gave up after a while what I can get [the nurse] to do, [be]cause that was taking too much of my cognitive load. (P19, ES = 4.50, COPD Case) "Kind of helps me talk to the team and where we're going and gives me a bit of time to think through where my next steps are...as we're getting ready for these steps." (P12, ES = 3.50, VFib / STEMI Case).
Subtheme 3C: Participants used pre-determined algorithms	Cognitive Resources	(Therapeutic Stage)-(Following Pre-Determined Algorithm) (Algorithms-Cognitive Demands) (Algorithms-Forgot) (Considering Algorithm) (Algorithm)-(Pre-Determined)-(Didn't Follow Through)		"...I have my predetermined or pre thought-out algorithm so ok, this is what COPD is, then I go on to my COPD box that I've learned in medical school and then… This is what I [want to] do, and I'm afraid that if I don't do it then [I'm going to] lose it. So, I [want to] get it out, start that treatment, and then reassess, and then I have time to think about things. I think it [kind of] has to do with this cognitive load. Once that's done then I can offload that information that I'm storing that's still in pieces instead of an easy-to-access box." (P12, ES = 3.50, COPD Case)

Themes and Subthemes	Example Categories	Example Codes	Low-Performers	High-Performers
Theme 4: Linear vs. Concurrent Approach *Subtheme 4A:* High-performers demonstrated a comfort level with moving back and forth between the orientation, diagnostic, and therapeutic stages of patient management.	Management Approach	(Patient Management)-(Concurrent) (Move Out of Stepwise Approach)		"So sick vs. not sick. When I walk in the room looking at the patient first is he awake, is he opening his eyes talking, kind [of] think moaning at the time looking at the monitor to see if I had, what they had already done for me. So, planning what, what I need to ask for immediately and then kind of I think looked distally at his limbs to see, you know, what's his colour like, is he perfused or not. And then immediately noticed that he was sick but not needing you know chest compressions or any immediate intervention" (P19H, ES= 4.50, GI Bleed) "He's hypotensive. Yeah. And here I'm just trying to assess, like neurologically what just happened," (P17, ES= 4.00, Hyperkalemia / Bradycardia Case)
Subtheme 4B: Low-performers tended to take a more linear approach to patient management.		(Patient Management)-(Linear) (Treatment Management)-(Linear Approach) (Therapeutic Stage)-(Stepwise Approach)	I've never actually paced anyone in real life so maybe I'm a bit hesitant to go that route...I think early med schools is always like, did you give any pain meds before you paced him? Oh my God, he's [going to] die, pace him. So, I think I'm still, you know, maybe a little hesitant on pacing" (P15, ES = 2.38, Hyperkalemia / Bradycardia Case)	

Anticipatory Behaviours

Anticipating potential problems and planning accordingly are key tasks in resuscitating a patient. Generally, high-performing residents were more strategic, considering multiple steps ahead of their current task. For example, one resident [P1H(Participant 1; high performer]) stated "So I'm anticipating for his melena stools and decreased [level of consciousness] that potentially he's bled out so much that he's quite hypovolemic and not perfusing his brain well. So, in anticipation of potentially needing a massive transfusion,

fluids etc. I want the best access on him I can with getting a central line." A low performing resident described another example of anticipation, "Even though it was PEA arrest so we we're going down the route of asking for the epi. But then, just wanting to have pads in case it, anything changed" [P4L (Participant 4; low performer)]. Anticipatory behaviours were seen across high and low-performing residents but were noted more often in high-performers as determined through coding frequencies.

Selective Attendance to Information

The ability to prioritize relevant stimuli, while appropriately deprioritizing less relevant stimuli is an essential skill during resuscitation cases. Generally, high-performers were able to discriminate between task relevant and task redundant information more often than the low performers based on coding frequencies. More specifically, when participants were given distractors (task redundant information), some struggled with prioritizing their therapeutic actions whereas others appropriately disregarded the non-essential information. An example of appropriate attendance to relevant stimuli was demonstrated by a high-performing resident when she was told about a high glucose level, "I'm more concerned about his airway at this point and getting interventions that way" (P17H). Similarly, another high-performing resident (P8H) quickly dismissed the incidental hyperglycemia distractor, "Weird. But like nothing that I needed to deal with imminently, so I was kind of like, I'll just store this in the back."

Another high performing participant described how she was simultaneously orienting herself and considering diagnostic possibilities, "So I was listening, looking at what was happening … I heard what I needed to hear in the first 5 seconds, I didn't need him to talk any further, but [the paramedic] was giving me all of the details, and in my head, I already said ok, this is probably [myocardial infarction]" (P12H).

In contrast, there were instances where the distractors caused residents to fixate. For example, a low-performing resident (P18L) stated, "… But then I got a bit distracted by the ECG and the bundle branch block and I thought maybe there was an anterior [myocardial infarction]."

Residents of all experience levels demonstrated the ability to selectively attend to information. However, the extent to which these behaviours were demonstrated tended to vary across residents, with higher performers exhibiting these behaviours more often.

Managing Cognitive Demands

Inherent to expert resuscitation is the ability to balance cognitive processing with the emergent needs of patients, family members, and team members. A lower-performing resident summed up this struggle by describing the "… demand in terms of mental resources, and [that] it can be sort of difficult to organize and compartmentalize that when there's a lot of stuff flying up in the air and a lot of decisions, then it's hard to prioritize them, I think" (P11L). In contrast, a high-performing resident described his experience with managing cognitive demands, "Yeah, so I can delegate the roles better because I shouldn't, like I don't [want to] task myself with drawing up medications and inserting IVs [be]cause of the cognitive load and [performing] CPR just seems like a terrible idea" (P1H).

Overall, we found that low-performing residents tended to focus on remembering algorithms associated with a given disease state resulting in potentially fewer cognitive resources to dedicate to emerging issues. For example a low-performing resident stated, "I want her on the monitors, I mean I'm also thinking I don't really know, it's not a huge deal I'm doing compressions right now, cause it's a pretty simple point of the algorithm right now" (P15L).

Linear vs. Concurrent Approach

Although all residents moved through the orientation (initial information gathering and actions upon entering the room), diagnostic, and therapeutic stages, low-performing residents demonstrated a linear approach while high-performing residents demonstrated a concurrent approach to patient management. The concurrent approach allowed high-performing residents to initiate treatments early while gathering information.

An example of the linear approach was described by a low performing trainee, “Yeah so I ended getting type and cross with all of my blood work because I knew that the patient needed blood. But then in my head I was waiting for those results to come back before I gave blood and completely spazzed on not giving him O positive or O negative in the meantime” (P2L). This resident approached the decision of when to transfuse based on the result of a blood test instead of ordering the test and simultaneously starting the transfusion with universal donor blood.

Another resident (P12H) stated, “This could still just be a calcium channel or beta blocker [overdose], and then I was thinking, either way, calcium is not dangerous. And I should’ve just given it empirically before I got any bloodwork back...but as I was thinking that I had already ordered a few things and going down my bradycardia algorithm. The way you learn it is atropine, pace, and then think about other things. And so, I think this is me trying to step out of the algorithm but going back to the algorithm anyways.” In reflecting on the case, this resident recognized the limitations of a linear approach.

Other high-performing residents demonstrated comfort moving between the stages of patient management. “And here I’ve decided, like here that it’s an unstable bradycardia. Blood pressure was awful, he’s quite brady[cardic], altered [level of consciousness]. Um, normal glucose. I didn’t have further information at this point to suggest that it was something else, so if not, then to improve this gentleman’s hemodynamics, I should just pace him” (P13H).

Limitations

This study has several limitations. Participants were grouped according to performance at one instant in time. We did not longitudinally follow trainees to determine how their cognitive processes evolved over time. As a result, we are unable to discuss how cognitive processes evolve at the individual physian level. Another limitation is that only one rater was used to score participants’ performance using the RAT. This being said, the QSAT (on which the RAT is based) has been shown to have an acceptable inter-rater reliability in the same simulation-based OSCE context (Dagnone et al., 2016; Hall et al., 2015), which suggests that using multiple raters would likely have produced similar results.

The traditional debriefing session (that preceded the study interview) was not standardized and it is possible that this session may have influenced participants' responses in the subsequent interview. Participants' responses in the interview may also have been biased by the Hawthorne effect and/or by their relationship with the interviewer. In addition, our previous research into the cognitive processes of physicians as well as our background in educational psychology likely coloured and influenced our qualitative analysis which was meant to be emergent. As a result, we may have had more difficulty recognizing themes alternate to what we were expecting to find. Despite this unavoidable bias, which should have made our data converge on these biases, we found differences between the cognitive processes of high and low performing participants. We believe that this adds further to the credibility of our findings.

Discussion

In this study, we found that the cognitive processes of high performing trainees differed from those of low performing trainees in a simulated resuscitation-based examination. High performers described more effective anticipatory behaviours, an improved ability to selectively attend to information, the ability to better manage their cognitive load, and exhibited a more concurrent (rather than linear) cognitive approach to patient management compared with low performing trainees.

These findings provide insight into the cognitive processes that underpin CRM skills within a simulated resuscitation context. By deriving these elements from within this context (as opposed to borrowing CRM principles from other fields, such as aviation) this may give medical educators a more authentic view into the way that medical simulation challenges trainees to think during a crisis.

The first derived theme was that of anticipatory behaviours. The ability to anticipate and plan for contingencies is a key skill in managing a complicated patient requiring resuscitation (Carne, Kennedy, M., & Gray, 2012). By considering multiple steps ahead and planning for potential failures, high performing trainees in this study consistently described this ability. As expected, low performers generally described less elaborate contingency plans. In both

the simulated and clinical context, it is usually clear when a trainee exhibits inadequate anticipation and his/her plan does not progress as expected. What is less clear to an outside observer is when the case progresses as the trainee expected and a lack of appropriate anticipation does not become apparent. The ability to delve into the thought processes of trainees as was done in this study may allow educators to better assess and teach these otherwise critical (but hidden) cognitive processes.

A second theme that was identified was selective attendance to information. Selective attendance to information refers to the ability of physicians to filter the innumerable stimuli of a resuscitation case and focus their attention on what is relevant, while appropriately de-prioritizing what is not. This phenomenon has been previously termed the *information reduction hypothesis* (Haider & Frensch, 1996). An improved ability to reduce extraneous information and focus on the most relevant environmental cues has been correlated with examination performance in another simulation-based study (A. Szulewski et al., 2019).

The third theme revolved around managing cognitive load. In a resuscitation case, appropriately dealing with cognitive demands is an important skill given the demands and pressures imposed on the team leader by these types of cases (Croskerry & Sinclar, 2001). In this study, high performers were more aware of their cognitive load and used a variety of strategies to manage cognitive demands, including capitalizing on shared mental models, utilizing algorithms, and delegating tasks to others.

Finally, higher performers in this study described a concurrent (as opposed to linear) approach to patient management. A traditional linear approach (typically taught to medical students) consists of performing a history, physical examination, ordering/analyzing tests, generating a diagnosis, and then initiating a treatment plan. This contrasts with the concurrent approach of simultaneous diagnostic and therapeutic actions that is characteristic of real-world patient care in the emergency department (A. Szulewski, Brindley, & van Merrienboer, 2017).

We hypothesize that the identified cognitive skills relate in different ways to the management of cognitive load that is seen with expertise development. With experience, learners gain the ability to create and automate new mental schemas (Sweller, Ayres, &

Kalyuga, 2011). This results in cognitive efficiency, effectively increasing available cognitive capacity, leading to better cognitive management, the ability to anticipate, utilize a concurrent approach, and selectively attend to information.

Several parallels can be drawn between these results, and those of a prior study conducted by our research team that focused on describing the cognitive processes underpinning CRM skills by expert trauma team leaders (TTLs) while leading real trauma resuscitations. Comparing the results of these two studies is informative. Two of the themes derived in the current study (*selective attendance to information* and *anticipatory behaviours*) also emerged in the TTL study (White et al., 2018). In addition, the theme of *managing cognitive demands* found in the current study aligns with the *cognitive load* subtheme of that manuscript. Importantly, these three parallel themes were more developed in the high performing group in the current study, which is consistent with expertise development theory (K. A. Ericsson, Prietula, & Cokely, 2007). Based on the snapshot in time that this study provides, it appears that the cognitive processes that underpin real-world CRM skills develop and become more expert-like as residents' performance improves in the simulation lab. The parallels between the cognitive processes seen in residents in the simulated test setting and those of expert TTLs in a real clinical setting (White et al., 2018), suggest that, to some degree, simulation provides the learning environment necessary for trainees to immerse themselves in resuscitation, which should positively impact their education (Hagiwara et al., 2016; Rehmann, Mitman, & Reynolds, 1995).

It is also important to discuss the differences between the themes derived here with those derived in real clinical settings. In real clinical settings, the transition to a concurrent patient management strategy was not observed (White et al., 2018). This is most likely because expert TTLs have developed an *expert blind spot* – they are so removed from the traditionally taught linear approach to patient management that they may be unaware that they have abandoned it (Nathan & Petrosino, 2003). Related to this idea is the concept of unconscious competence in the context of expertise development, where task performance becomes *second nature* for experienced practioners (Burch, 1970). Further, the themes of logistical awareness and visual gaze behaviours were also absent in the current study. This is most likely related to differences between simulated and real-life contexts. The awareness of time and prioritization of real-world tasks characteristic of logistical awareness, as well as

visualization/recognition of individuals and equipment that constitute directed visual gaze, are artificial in a simulated environment where delays do not exist, and equipment and personnel are predictably available and reliable.

This discrepancy is worthy of further analysis. If the goal of resuscitation-based simulation examinations is to assess learners' abilities to perform the tasks required in real clinical practice, then as simulation educators we should aim to create scenarios that incorporate realistic time constraints/delays that better simulate the messiness of the real world. Doing so has the potential to further improve the functional task alignment of simulation.

Additional attention should focus on the process of debriefing learners following both simulated and real-life experiences. Cheng and colleagues recently summarized educational strategies to improve resuscitation outcomes. A key approach that was identified was focusing on learner debriefing following resuscitation (Cheng et al., 2018). The method of debriefing used in the current study serves as one novel example of how this can be implemented successfully.

By providing insight into the cognitive processes of residents managing simulated patients during examinations, this study supports the current practice of using simulation in residency training. At the same time, the study suggests that simulation training offers an imperfect representation of the messiness of the real clinical environment. Future studies are needed to test whether, as educators, we can improve simulation teaching and assessment to better prepare and arm our learners with the CRM skills required for real-world practice.

Appendix A

Consolidated criteria for reporting qualitative studies (COREQ): 32-item checklist

Developed from:

Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*. 2007. Volume 19, Number 6: pp. 349 – 357

No. Item	Guide questions/description	Reported on Page #
Domain 1: Research team and reflexivity		
Personal Characteristics		
1. Interviewer/facilitator	Which author/s conducted the interview or focus group?	Page 7
2. Credentials	What were the researcher's credentials? E.g. PhD, MD	Page 10
3. Occupation	What was their occupation at the time of the study?	Page 10
4. Gender	Was the researcher male or female?	Page 10
5. Experience and training	What experience or training did the researcher have?	Page 10
Relationship with participants		
6. Relationship established	Was a relationship established prior to study commencement?	Page 10 .
7. Participant knowledge of the interviewer	What did the participants know about the researcher? e.g. personal goals, reasons for doing the research	Page 10
8. Interviewer characteristics	What characteristics were reported about the interviewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic	Page 10

No. Item	Guide questions/description	Reported on Page #
Domain 2: study design		
Theoretical framework		
9. Methodological orientation and Theory	What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis	Page 4
Participant selection		
10. Sampling	How were participants selected? e.g. purposive, convenience, consecutive, snowball	Page 5
11. Method of approach	How were participants approached? e.g. face-to-face, telephone, mail, email	Page 5
12. Sample size	How many participants were in the study?	Page 11
13. Non-participation	How many people refused to participate or dropped out? Reasons?	Page 11
Setting		
14. Setting of data collection	Where was the data collected? e.g. home, clinic, workplace	Page 5 .
15. Presence of non-participants	Was anyone else present besides the participants and researchers?	No
16. Description of sample	What are the important characteristics of the sample? e.g. demographic data, date	Pages 5 and 11
Data collection		
17. Interview guide	Were questions, prompts, guides provided by the authors? Was it pilot tested?	Page 7
18. Repeat interviews	Were repeat interviews carried out? If yes, how many?	No
19. Audio/visual recording	Did the research use audio or visual recording to collect the data?	Page 6
20. Field notes	Were field notes made during and/or after the interview or focus group?	No
21. Duration	What was the duration of the interviews or focus group?	Pages 6, 7
22. Data saturation	Was data saturation discussed?	Page 9
23. Transcripts returned	Were transcripts returned to participants for comment and/or correction?	No

No. Item	Guide questions/description	Reported on Page #
Domain 3: analysis and findings		
Data analysis		
24. Number of data coders	How many data coders coded the data?	Page 9
25. Description of the coding tree	Did authors provide a description of the coding tree?	Pages 23-27
26. Derivation of themes	Were themes identified in advance or derived from the data?	Pages 8-9
27. Software	What software, if applicable, was used to manage the data?	Page 9
28. Participant checking	Did participants provide feedback on the findings?	No
Reporting		
29. Quotations presented	Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? e.g. participant number	Pages 11-16, 24-27
30. Data and findings consistent	Was there consistency between the data presented and the findings?	Pages 10-14, 24-27
31. Clarity of major themes	Were major themes clearly presented in the findings?	Pages 10-14, 24-27
32. Clarity of minor themes	Is there a description of diverse cases or discussion of minor themes?	Pages 10-14, 24-27

Appendix B

Example clinical vignette that participants read prior to starting the case

STATION 1

A patient has been brought into a resuscitation room by EMS with cardiac arrest.

EMS Patch: 65-year-old male with one hour of chest pain. En route to the emergency department, he had a sudden loss vitals.

An ED nurse and a paramedic are currently available to help you with the assessment and care of this patient.

You will have 8 minutes to complete this station.

The nurse will let you know when to enter the room.

Appendix C

Emergency Medicine Resuscitation Assessment Tool

Trainee Identification: _____ Date of Assessment: _____

Assessed by: _____

EPA = Resuscitate and manage the care of the critically ill medical/surgical patient

Clinical Context (Case) = _____

Primary Assessment				
• *Ensures monitors are applied & vital signs obtained (incl glucose + Temp)* • *Establishes appropriate vascular access* • *Conducts a focused assessment of airway & breathing*		• *Assesses level of consciousness/disability* • *Simultaneously performs initial diagnostic & initial therapeutic/resuscitative actions* • *Allocates & utilizes resources appropriately*		
1 Observation Only	**2** Direct Supervision	**3** Indirect Supervision	**4** Independent	**5** Supervision of Trainees

Diagnostic Actions				
• *Performs a targeted history & physical exam* • *Exposes the patient appropriately to complete exam* • *Orders appropriate blood work*		• *Performs rhythm analysis/ECG as indicated* • *Performs targeted point of care ultrasound as indicated* • *Orders appropriate imaging*		
1 Observation Only	**2** Direct Supervision	**3** Indirect Supervision	**4** Independent	**5** Supervision of Trainees

Therapeutic Actions				
• *Prioritizes critical or time sensitive therapies* • *Performs/directs necessary resuscitative maneuvers* • *Manages airway & ventilator support as needed*		• *Orders IV fluids or blood products as appropriate* • *Orders appropriate medications as required* • *Coordinates disposition & specialist involvement*		
1 Observation Only	**2** Direct Supervision	**3** Indirect Supervision	**4** Independent	**5** Supervision of Trainees

Communication				
• *Uses clear, directed, closed loop communication* • *Clearly assigns & articulates leadership* • *Shares mental model & verbalizes priorities*		• *Solicits opinion from team members, experts, & consultants as needed* • *Involves patient & family in decision-making* • *Prepares & debriefs team as time permits*		
1 Observation Only	**2** Direct Supervision	**3** Indirect Supervision	**4** Independent	**5** Supervision of Trainees
ENTRUSTMENT DECISION				
1 *Observation Only*	*2* *Direct Supervision*	*3* *Indirect Supervision*	*4* *Independent Performance*	*5* *Supervision of Trainees*

SPECIFIC RATIONALE:

Appendix D

Additional quotations organized by themes and sub-themes

Themes and Subthemes	Example Quotations
Theme 1: Anticipatory Behaviours *Subtheme 1A:* Generally, high-performers were more strategic, considering multiple steps ahead of their current task.	"I was doing it so I could titrate, I expected his blood pressure to fall and I wanted to catch it." (P20, ES = 2.00, low performer, VFib/STEMI case) "These were going to be our steps if we couldn't- Like first was direct, then bougie, then glideoscope so I'd already thought about the glidescope earlier." (P7, ES = 3.13, high performer, GI Bleed case)
Theme 2: Selective Attendance to Information *Subtheme 2A:* Generally, high-performers were able to discriminate between task relevant and task redundant information more often than the low performers.	"I didn't really think that it was going to change anything that we were going to do in the room and it's just adding extra noise for them to think about." (P5, ES = 3.25, high performer, VFib/Stemi case) "And I was like well is there anything else that I need to worry about in this guy and then I realized very quick- I hope very quickly that he was just yammering on and nothing he was really saying was going to help me and that I needed his help with dialysis eventually so he should just come here, rather than talking on the phone." (P3. ES = 3.00, high performer, Hyperkalemia / Bradycardia case) "And then I'm thinking like when can I stop him and eventually I'm just like that's enough" (P27, ES = 2.00, low performer, VFib/STEMI case) "I said this a critical low oxygen on a VBG so we're going to ignore it." (P5, ES = 3.25, high performer, VFib/Stemi case) "so I just wanted to shut him down cause, you know, this is, he's arrested, I need to know whether he's shockable, I wanna give him a shock." (P13, ES = 3.50, high performer, VFib/STEMI case) "Yeah and it was. Like ok, he's being resuscitated now, you can come and see him if you want. (laughs)…That's all I had, if you want more, you need to come down and see him." (P16, ES = 3.00, high performer, Hyperkalemia / Bradycardia case) "Getting a sugar distractor. A guy with CP. [..] wasn't too worried about. I'm kind of, you can see I kind of dismissed her pretty quickly because I started looking at the monitor, it was just fine." (P15, ES = 2.38, low performer, Seizure/Meningitis/Hypoglycemia case)

Themes and Subthemes	Example Quotations
Theme 3: Managing Cognitive Load *Subtheme 3A*: Low-performers tended to have fewer cognitive resources to dedicate to emerging issues. *Subtheme 3B*: High-performers tended to utilize a variety of strategies to manage their cognitive demands. *Subtheme 3C:* Participants used pre-determined algorithms	"And then I was too overloaded because I wasn't necessarily going to use the information right then and there" (P24, ES = 1.50, Hyperkalemia / Bradycardia case) "I felt a little bit cognitive overloaded with that…." (P14, ES = 2.50, low performer, GI Bleed case) "That's the thing I was looking for the whole time, that there's something that was going on in this guy, that obviously had some medical history, because he's coming from a group home when he's 50. So I think I hung on to what's going on, because there's an answer, and I delayed my generic supportive management, and then once that actually came, I started the supportive management and I had overloaded myself." (P20, ES = 2.00, low performer, Hyperkalemia / Bradycardia case) "Gave me a chance to think about it. So actually right here I'm writing my 5 H's and 5 T's down before I go in." (P10, ES = 3.50, high performer, VFib/STEMI case) "To be honest I had no interest in what that other person was saying, because their patient in my mind couldn't have been more sick than mine or if they were, there was nothing at the moment I could do to help them. I probably looking back could've said I'm busy right now can you keep them on the line or I'm busy right now have them call me back in 10 minutes but at the moment it was, I don't want to deal with this right now, I'm going to have them remove it entirely from my mind and move ahead with my intubation. (P19, ES = 4.50, high performer, COPD case) "Yeah. I think we talk about this in orals. And it's just about the packages. To offload that work load. So. The vitals were pretty easy to kind of package off." (P27, ES = 2.00, low performer, Hyperkalemia / Bradycardia case) **"**Because that one [arrest algorithm] really doesn't require a lot of thought." (P11, ES = 2.00, low performer, GI Bleed case) "So here I'm just sort of going back through my mind thinking what have we got going, what's running, what am I missing. You know. Thinking through sort of H's and T's of PEA." (P6, ES = 2.63, low performer, GI Bleed case) "Yeah, so there I'm thinking, you know my two pathways on ACLs; one is medical management. Need somebody who looks well but who has a slow heart rate versus altered pain, anything that's sort of extra symptoms, then going down the unstable pathway." (P6, ES = 2.63, low performer, Hyperkalemia / Bradycardia case)
Theme 4: Linear vs. Concurrent Approach *Subtheme 4A:* High-performers demonstrated a comfort level with moving back and forth between the orientation, diagnostic, and therapeutic stages of patient management. *Subtheme 4B:* Low-performers tended to take a more linear approach to patient management.	"Just thinking that she's giving me this story where I've already sort of thought that this guy might have a quintessential GI bleed, um and then just wanna see how he's doing sort of as a quick assessment from his ABC perspective while she tells me" (P11, ES = 2.00, low performer, GI Bleed case) "Still on the differential, uh and I think I gave like half dose of calcium, just because for cardiac stability I thought of all the things I would give him, it would be calcium stabilizing myocardium but I didn't want to give him 2 grams because I was also worried, he's a dialysis patient, he could be hyperkalemic right now." (P16, ES = 3.00, high performer, Hyperkalemia / Bradycardia case) "So then I'm thinking about like what the cause of the arrest was, and then just making sure that you're maximizing both post-resuscitation care, or post-ROSC care." (P17, ES = 4.00, VFib / STEMI case) "So then I look at this and I again I couldn't really see what the p waves were doing. But I was fairly certain like it wasn't a first or second degree block so I just decided that ok we need to pace." (P3, ES = 3.00, high performer, Hyperkalemia / Bradycardia case)

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi:10.1191/1478088706qp063oa

Burch, N. (1970). *The four stages for learning any new skill*. Solana Beach, CA: Gordon Training International.

Carne, B., Kennedy, M., & Gray, T. (2012). Crisis resource management in emergency medicine *Emergency Medicine Australasia, 24*(1), 7-13. doi:10.1111/j.1742-6723.2011.01495.x

Cheng, A., Nadkarni, V. M., Mancini, M., Hunt, E. A., Sinz, E. H., Merchant, R. M., . . . Bhanji, F. (2018). Resuscitation Education Science: Educational Strategies to Improve Outcomes From Cardiac Arrest. *Circulation, 138*(6), e82-e122. doi:10.1161/CIR.0000000000000583

Cook, D. A., Andersen, D. K., Combes, J. R., Feldman, D. L., & Sachdeva, A. K. . (2018). The value proposition of simulation-based education. *Surgery, 163*(4), 944-949.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches*. Thousand Oaks, CA: Sage.

Croskerry, P., & Sinclar, D. (2001). Emergency medicine: A practice prone to error? . *Canadian Journal of Emergency Medicine, 3*(04), 271-276.

Dagnone, D., Hall, A., Sebok-Syer, S., Klinger, D., Woolfrey, K., Davison, C., ... (2016). Competency-based simulation assessment of resuscitation skills in emergency medicine postgraduate trainees – a Canadian multi-centred study. *Canadian Medical Education Journal, 7*(1), e57-e67.

Davidsen, A. (2013). Phenomenological Approaches in Psychology and Health Sciences. *Qualitative Research in Psychology, 10*(3), 318-339.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. *The Cambridge handbook of expertise and expert performance*, 223-241.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7/8), 114-121.

Gegenfurtner, A., Kok, E., van Geel, K., de Bruin, A., Jarodzka, H., Szulewski, A., & van Merriënboer, J. J. (2017). The challenges of studying visual expertise in medical image diagnosis. *Medical Education, 51*(1), 97-104.

Giorgi, A. (2009). *The descriptive phenomenological method in psychology: A modified Husserlian approach*. Pittsburgh, PA.: Duquesne University Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Hagel, C. M., Hall, A. K., & Dagnone, J. D. (2016). Queen's University Emergency Medicine Simulation OSCE: an Advance in Competency-Based Assessment. *Canadian Journal of Emergency Medicine, 18*(3), 230-233. doi:10.1017/cem.2015.34

Hagiwara, M. A., Backlund, P., Maurin, H., Soderholm, H. M., Lundberg, L., Lebram, M., & Engstrom, H. (2016). Measuring participants' immersion in healthcare simulation the development of an instrument. *Advances in Simulation, 1*(17), 1-9.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30*(3), 304-337.

Hall, A. K., Dagnone, J. D., Lacroix, L., Pickett, W., & Klinger, D. A. (2015). Queen's Simulation Assessment Tool: development and validation of an assessment tool for resuscitation Objective Structured Clinical Examination Stations in emergency medicine. *Simulation in Healthcare, 10*, 98-105.

Helmreich, R. L. (2000). On error management: lessons from aviation. *British Medical Journal, 320*(7237), 781.

Hicks, C. M., Bandiera, G. W., & Denny, C. J. . (2008). Building a Simulation-based Crisis Resource Management Course for Emergency Medicine, Phase 1: Results from an Interdisciplinary Needs Assessment Survey. *Academic Emergency Medicine, 15*(11), 1136-1143. doi:10.1097/01.CCM.0000229877.45125.CC

Johnson, G., & Reynard, K. (1994). Assessment of an objective structured clinical examination (OSCE) for undergraduate students in accident and emergency medicine. *Journal of Accident & Emergency Medicine, 11*(4), 223-226.

Kim, J., Neilipovitz, D., Cardinal, P., Chiu, M., & Clinch, J. (2006). A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical Care Medicine, 34*(8), 2167-2174.

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education, 51*(1), 114-122.

McIntyre, R. M., & Salas, E. (1995). Team performance in complex environments: What we have learned so far. *Team effectiveness and decision making in organizations*, 9-45.

Mills, B. W., Carter, O. B., Rudd, C. J., Claxton, L. A., Ross, N. P., & Strobel, N. A. (2016). Effects of Low- Versus High-Fidelity Simulations on the Cognitive Burden and Performance of Entry-Level Paramedicine Students: A Mixed-Methods Comparison Trial Using Eye-Tracking, Continuous Heart Rate, Difficulty Rating Scales, Video Observation and Interviews. *Simulation in Healthcare, 11*(1), 10-18. doi:10.1097/SIH.0000000000000119

Nathan, M. J., & Petrosino, A. (2003). Expert Blind Spot Among Preservice Teachers. *American Educational Research Journal, 40*(4), 905-928.

Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2017b). A Novel Approach to Study Medical Decision Making in the Clinical Setting: The "Own - point - of - view" Perspective. *Academic Emergency Medicine, 24*(7), 785-795. doi:10.1111/acem.13209

Rehmann, A., Mitman, R., & Reynolds, M. (1995). *A Handbook of Flight Simulation Fidelity Requirements for Human Factors Research.* Retrieved from OH, USA

Reznek, M., Smith-Coggins, R., Howard, S., Kiran, K., Harter, P., Sowb, Y., Gaba, D., Krummel, T. (2003). Emergency Medicine Crisis Resource Management (EMCRM): Pilot study of a simulation-based crisis management course for emergency medicine. *Academic Emergency Medicine, 10*(4), 386-389.

Salas, E., Rosen, M. A., & King, H. (2007). Managing teams managing crises: principles of teamwork to improve patient safety in the emergency room and beyond. *Theoretical Issues in Ergonomics Science, 8*(5), 381-394.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A., & Pappas, J. (2013). Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department. *Annals of Emergency Medicine, 61*(1), 96-109. doi:10.1016/j.annemergmed.2012.08.034

Schull, M. J., Ferris, L. E., Tu, J. V., Hux, J. E., & Redelmeier, D. A. (2001). Problems for clinical judgement: Thinking clearly in an emergency. *Canadian Medical Association Journal, 164*(8), 1170-1175.

Strauss, A. L., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage Publications, Inc.

Sundar, E., Sundar, S., Pawlowski, J., Blum, R., Feinstein, D., & Pratt, S. (2007). Crew resource management and team training. *Anesthesiology Clinics, 25*(2), 283-300.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Amassing Information: The Information Store Principle*. New York, NY: Springer.

Szulewski, A., Braund, H., Egan, R., Hall, A., Dagnone, D., Gegenfurtner, A., Merrienboer, J. V. (2018). In the eyes of the learner: Eye-tracking augmented debriefing in simulation. *Journal of Graduate Medical Education., 10*(3), 340-341. doi:10.4300/JGME-D-17-00827.1

Szulewski, A., Brindley, P. G., & van Merrienboer, J. J. G. (2017). Decision making in acute care medicine. In P. G. Brindley & P. Cardinal (Eds.), *Optimizing crisis resource management to improve patient safety and team performance: A handbook for all acute care health professionals*. Canada: Practice, performance and innovation unit of the Royal College of Physicians and Surgeons of Canada.

Szulewski, A., Egan, R., Gegenfurtner, A., Howes, D., Dashi, G., McGraw, N. C. J., . . . van Merrienboer, J. J. G. (2019). A new way to look at simulation-based assessment: the relationship between gaze-tracking and exam performance. *Canadian Journal of Emergency Medicine, 21*(1), 129-137. doi:10.1017/cem.2018.391

Ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *J Grad Med Educ 5*, 157-158.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. . *International Journal for Quality in Health Care, 19*(6), 349-357.

Weersink, K., Hall, A. K., Rich, J., Szulewski, A., & Dagnone, D. (2018). *How does resuscitation performance in the simulation setting relate to performance in the real world? A direct comparison of entrustment scoring of emergency medicine residents.* Paper presented at the International Conference on Residency Education, Halifax, NS, Canada.

White, M. R., Braund, H., Howes, D., Egan, R., Gegenfurtner, A., van Merrienboer, J. J., & Szulewski, A. (2018). Getting Inside the Expert's Head: An Analysis of Physician Cognitive Processes During Trauma Resuscitations. *Annals of Emergency Medicine, 72*(3), 289-298. doi:10.1016/j.annemergmed.2018.03.005

Chapter 7

General Discussion

People have been interested in knowing what makes certain individuals *experts* for centuries. Consider a concert pianist, chess grandmaster or professional athlete – more than just executing the tasks required of them, domain experts can make it look easy. This dissertation has examined the development of expertise, in detail, in one particular domain – that of resuscitation medicine. Comprised of declarative knowledge, procedural ability and crisis resource management skills, resuscitation medicine is the embodiment of a complex skill in medicine – where physicians are tasked with caring for the acutely unwell during medical crises. Grounded in educational psychology, this dissertation has strived to answer the following overarching research question that was introduced in Chapter 1:

> *What changes in cognitive load and processing occur in physicians as they develop expertise within the domain of resuscitation medicine? How can these changes be measured as physicians progress along the expertise continuum?*

As medical educators, it is incumbent upon us to help move our trainees forward along this novice-expert continuum – but to do this effectively, we must first understand how experts think and perform. In an effort to accomplish this goal, this dissertation has focused on the following research questions which are answered in Chapters 2 – 6:

1. What is the relationship of cognitive load, as measured by pupillometry, to level of experience in the context of a traditional knowledge-based resuscitation medicine examination?
2. In the context of a resuscitation medicine examination, what is the validity of using a physiologic measure of cognitive load (pupillometry) and a psychometric one (Paas, 1992) as markers of cognitive load among physicians with different levels of experience?
3. In resuscitation-based simulation objective structured clinical examinations (OSCEs), how (and to what extent) are particular gaze patterns of residents associated with exam performance? How do these gaze patterns vary across scenarios?
4. What are the cognitive processes of expert physicians while leading actual trauma resuscitations?
5. What are the cognitive processes of medical trainees in simulation-based resuscitation examinations? How do these cognitive processes vary with examination performance?

This general discussion begins with a description of the main findings as they relate to each of the research questions and to the overarching research question. Next, theoretical contributions of the work are discussed, followed by a section on limitations. The chapter closes with a general conclusion.

Main findings

The main findings of this dissertation can be divided into two main sections: cognitive load measurement and cognitive process measurement in the realm of resuscitation medicine expertise development. Chapters 2 and 3 of this dissertation focus on cognitive load measurement; while Chapters 4, 5 and 6 deal with cognitive process measurement. An overview of how the chapters fit within this structure is shown in Figure 1.

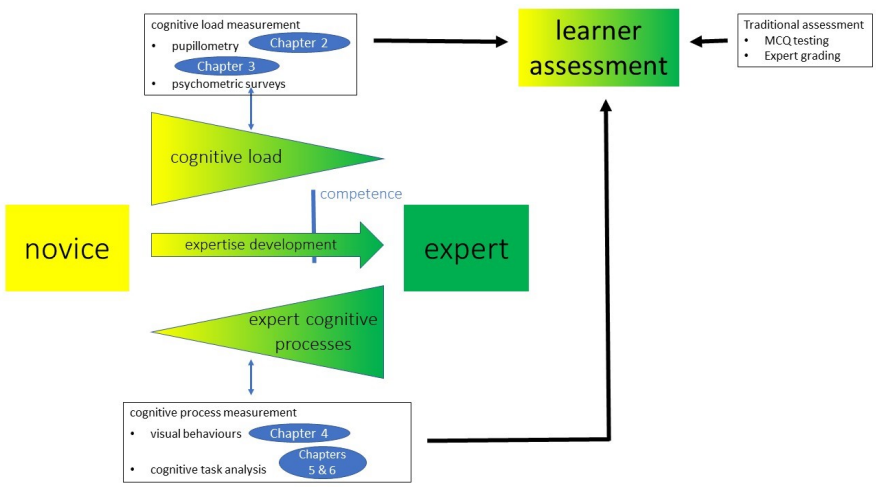


Figure 1: Theoretical framework for the analysis of the development and assessment of expertise in resuscitation medicine. Also shown is where each of the studies conducted fits within this framework.

Research question 1: *What is the relationship of cognitive load, as measured by pupillometry, to level of experience in the context of a traditional knowledge-based resuscitation medicine examination?*

In its simplest terms, cognitive load refers to the effort being made in processing information in working memory (Sweller, van Merrienboer, & Paas, 1998). There are three main techniques by which cognitive load can be measured empirically: analysis of physiologic variables, responses to psychometric surveys and dual-task methodology (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). Pupillometry is one type of physiologic variable that has been used to measure cognitive load for decades. It is based upon the predictable dilation of humans' pupils with increasing cognitive demands and the corresponding pupillary constriction that occurs when cognitive demands decrease due to sympathetic nervous system activity (Laeng, Sirois, & Gredebäck, 2012). [1] In Chapter 2, pupillometry was used to study novice and trained physicians as they answered arithmetic, general knowledge, and resuscitation-medicine-based questions in a multiple choice examination format. We found that cognitive load in all participants increased as arithmetic questions became more difficult. There were no changes in measured cognitive load between the two groups when answering general knowldege questions. However, when answering domain-specific (resuscitation-medicine focused) questions, novices experienced a significantly increased cognitive load compared to trained physicians. Moreover, when examining the subgroup of questions answered correctly by both groups (where both groups would have scored the same grade on traditional assessment metrics), trained physians experienced a significantly decreased cognitive load compared to novices. The results suggest that the organization of knowledge in memory of novices in resuscitation medicine is indeed different than that of those with experience within this domain. In addition, these findings suggest that there may be more objective ways of categorizing expertise related to resuscitation medicine over and beyond traditional notions of examination performance based on a percentage of correct answers. The study suggests that cognitive load

[1] In Appendix A, we review the evidence surrounding pupillometry as a surrogate marker for cognitive load (Szulewski, Kelton, & Howes, 2017).

measurement can add important objective information to assessment data that has not been previously available to medical educators.

Research question 2: *In the context of a resuscitation medicine examination, what is the validity of using a physiologic measure of cognitive load (pupillometry) and a psychometric one (Paas, 1992) as markers of cognitive load among physicians with different levels of experience?*

Though it seems likely that the construct that was measured by pupillometry in Chapter 2 was in fact cognitive load, questions could be raised about the construct validity of using pupillometry as a surrogate marker in this context. Adding to the validity argument that pupillometry measures the construct of cognitive load in a test-taking environment, Chapter 3 provides evidence grounded in Messick's validity framework (Cook & Beckman, 2006). In this study, we compared two purported measures of cognitive load head-to-head: pupillometry and psychometric surveying – both in a resuscitation medicine as well as an arithmetic examination format. In this study, we found that both tools reflected the expected changes in cognitive load based on question type, difficulty, participant accuracy, and level of training. Further, there was a strong positive correlation between the two measurement tools. Grounded in Messick's validity framework (Cook & Beckman, 2006), this study provided construct validity in the form of internal structure, response process and relations to other variables.

Taken in context with the findings from Chapter 2 as well as Appendix A, Chapter 3 provides evidence to suggest that physician cognitive load can indeed be measured in a controlled examination environment. These findings are interesting when considered from a medical education standpoint. Knowing that there are predictable differences in the amount of cognitive load that physicians with differing levels of expertise are required to dedicate to a task opens avenues for future research to better understand cognitive architecture and possibly to use cognitive load as a guide to tailor the level of instructional support that is required for a particular learner.

Research question 3: *In resuscitation-based simulation OSCE's, how (and to what extent) are particular gaze patterns of residents associated with exam performance? How do these gaze patterns vary across scenarios?*

The second half of this dissertation moved from cognitive load measurement to cognitive process measurement. In this dissertation, cognitive processes were described by analyzing visual behaviours as well as performing post-hoc cognitive task analyses on physicians. Eye-tracking is a useful technology that allows researchers to objectively quantify vision and analyze visual behaviours. Because any visual stimulus that is attended to and processed must first be seen, eye tracking provides a window into the viewer's mind (Kok & Jarodzka, 2017). It has been established that in the medical specialties that rely on visual diagnosis (like radiology), the visual patterns of novices and experts are different (Gegenfurtner et al., 2017). The goal of Chapter 4 of this dissertation was to determine whether visual differences existed in the highly complex field of resuscitation medicine – a discipline that is not generally considered a visual one. In this study, we found that residents' gaze patterns were correlated with objective performance in the context of a resuscitation-based examination in a high-fidelity simulation laboratory. Individuals who performed better had an improved ability to deprioritize task-redundant information appropriately and selectively process task-relevant information. Those who performed poorly did not have this same ability for selective prioritization. Of note, task-relevant (vs. task-redundant) stimuli were found to differ between various cases. Take, for example, the electrocardiogram. Though critical to analyze in a case of a patient presenting with chest pain, spending time analyzing an electrocardiogram in a trauma patient would be a waste of valuable time that would be better spent actively resuscitating the patient. Physicians with more experience tended to have the ability to make this distinction more readily and focused on the appropriate tasks to optimize patient care. This phenomenon has been previously described in the expertise literature in non-medical contexts. Haider and Frensch (1999) called this the information reduction hypothesis and suggested that experts' ability to limit information at a perceptual (as opposed to conceptual) level allows them to be more efficient when interacting with their environments.

The parallels that can be drawn between the complex environment of a resuscitation to the more predictable environment of the visual specialties are important for medical education.

Knowing the proportion of time that a physician focuses on task relevant vs. redundant stimuli may give educators additional data to determine where a learner lies on the novice-expert continuum.

Research question 4: *What are the cognitive processes of expert physicians while leading actual trauma resuscitations?*

Continuing the analysis into cognitive process measurement, the last sections of this dissertation used cognitive task analysis techniques to delve into the minds of resuscitation medicine providers. [2]

Chapter 5 looked at the broad scope of cognitive processes as expert trauma team leaders managed actual trauma patients. In this study, experts wore eye-trackers that recorded the clinical scene from a first-person perspective. The resultant video was then viewed by the participants and used to augment traditional cognitive task analysis in an effort to get inside the heads of the expert group. Using qualitative methodology, we found that experts in trauma resuscitation shared a number of cognitive processes with one another: logistic awareness, managing uncertainty, directed visual gaze, selective attendance to information and anticipation of pitfalls. The derivation of these cognitive processes from expert physicians doing real clinical work represents the first foray into the study of crisis resource management skills during real medical crises. Armed with this better understanding of resuscitation medicine expertise, we have been able to uncover some of the elements of cognition that are usually considered tacit and difficult for experts themselves to describe. Having identified these cognitive processes may allow medical educators to tailor both their teaching and assessment strategies appropriately.

[2] Appendix B outlines theory surrounding a central component of physician cognitive processes – that of decision-making during medical crises. It contextualizes decision-making theory within crisis resource management and resuscitation medicine (Szulewski, Brindley, & Van Merrienboer, 2017).

Research question 5: *What are the cognitive processes of medical trainees in simulation-based resuscitation examinations? How do these cognitive processes vary with examination performance?*

Having elucidated some of the cognitive processes of expert physicians managing real resuscitation cases, we were interested in determining whether medical trainees exhibited similar cognitive processes when managing simulated cases. In Chapter 6, we examined the cognitive processes of emergency medicine trainees with a cognitive task analysis approach augmented by viewing the residents' own first-person video as recorded by the eye-tracker. Resident performance was also separately scored by a blinded external reviewer. Higher performing residents were better able to anticipate, selectively attend to relevant information, manage cognitive demands, and took a concurrent (as opposed to linear) approach to managing the patient. The results provide new insights into the cognitive processes of residents and also emphasize that the cognitive processes of high performing residents approach those of physicians in clinical practice (in contrast to those of lower performing residents). [3] Taken together, these findings add evidence that simulation training of resuscitation medicine skills provides at least some of the psychological fidelity required for learners as they progress along the novice-expert continuum.

Overarching research question: *What changes in cognitive load and processing occur in physicians as they develop expertise within the domain of resuscitation medicine? How can these changes be measured as physicians progress along the expertise continuum?*

Taken together, Chapters 2 – 6 answer the overarching research question of this dissertation. Focusing on the cognitive load measurement piece, the evidence shows that there is a measurable decrease in the amount of cognitive load that physicians with increasing levels of resuscitation medicine experience when they are required to complete a

[3] Appendix C describes how the novel technique of cued retrospective debriefing augmented by eye tracking used in Chapter 6 allowed learners to critique their responses to specific situational cues more accurately and to identify new insights into their own performance that would have been otherwise missed (Szulewski et al., 2018).

domain-specific task. Pupillometry and psychometric surveying were both found to be useful techniques to measure this. From a cognitive process measurement perspective, it appears that cognitive processes change and become more expert-like as learners progress along the expertise continuum based on studying visual behaviours as well as with cognitive task analysis techniques. As expertise in resuscitation medicine progresses, individuals are better at focusing on what's relevant in their environments and organizing their thoughts towards managing a medical crisis.

Though known to be true in other fields, these findings are novel in the domain of resuscitation medicine and medical education. These data suggest that medical educators may be able to move beyond traditional assessment modalities when measuring where individual trainees lie on the expertise continuum. Armed with this new information, educators could also potentially better guide and more individually tailor educational strategies for individual trainees because it is known that the effectiveness of particular strategies is highly dependent on the level of expertise of the individual learner. The information about cognitive load and cognitive processes provided by the novel application of these tools is not meant to replace traditional assessment, but it can add important objective information that has not been previously available to medical educators.

Theoretical implications and suggestions for future research

Towards a richer understanding of expertise development in resuscitation medicine

A central challenge in understanding expertise in any domain is that experts often find it difficult to accurately explain what makes them experts (Ericsson, 2018). Part of the reason is because, over time, many of the cognitive processes that underlie expert decision-making become tacit, automatic and may not even be viewed as decisions to the experts at all (Klein, Calderwood, & Clinton-Cirocco, 1986). In medicine in particular, expertise is difficult to study because a "gold standard" of performance is difficult to define (Ericsson, 2004).

These phenomena hold true across medical specialties, notably in the domain of emergency (and resuscitation) medicine despite there being a wide gap in the decision making and cognitive processes of novices and experienced providers (Schubert, Denmark, Crandall,

Grome, & Pappas, 2013). In this dissertation's introductory chapter, we suggested that the high-stakes nature of resuscitation medicine, its characteristic decision-making-under-pressure, and the ubiquity of acute illness across many medical specialties warranted taking a deep dive into expertise development in this field. Throughout the subsequent chapters, we have described this deep dive, grounded in empirical research.

By moving beyond traditional notions of performance, we sought to *get under the hood* of what it means to move along the novice-expert continuum in resuscitation. The underlying hypothesis of our work was that as physicians move along this continuum, their cognitive load would decrease for particular tasks and their cognitive processes would become more expert-like.

The results of the studies presented here support this framework. In Chapters 2 and 3 we provided evidence that experienced physicians expended less cognitive load when answering questions aimed at domain-specific declarative knowledge than more novice physicians. This is likely related to the changes in the way that knowledge is organized in memory in physicians as they progress along the novice-expert continuum (Van Merrienboer & Sweller, 2005). The resultant increase in remaining working memory capacity is important when these results are extrapolated to the clinical world. If more experienced physicians expend less cognitive load (i.e. mental effort) remembering, for example, the steps of an algorithm or the doses of medications, then they should have a corresponding increase in available working memory resources for other important tasks required of them. Because there are always numerous tasks to accomplish in a complicated resuscitation case, having available working memory capacity over and beyond what's required at a given time is an advantage in the clinical world. This extra capacity allows the experienced physician leader to take more time with family, managing multiple simultaneous patients, etc. Therefore, we should strive to measure extra capacity as trainees progress along the novice-expert continuum. Future research could help to delineate what individuals actually do with this increased cognitive capacity. A mixed methods approach could be utilized to help answer this question. Physicians of various levels of expertise in resuscitation medicine could be interviewed about their own cognitive processes (whether in the simulation lab while learning or in the clinical environment while working) that aren't directly related to task performance and these results stratified

according to additional extra working memory capacity (as quantified by 1 – measured cognitive load).

Chapters 4, 5 and 6 of this dissertation focused on how the cognitive processes of physicians change as they progress along the novice-expert continuum. Grounded in an analysis of visual behaviours and analogous to previous work by Haider and Frensch (1996), Chapter 4 showed that more experienced physicians are better able to filter their surroundings than novices. By focusing on what's important and appropriately de-prioritizing what's not, high performers were found to take in less visual task-redundant information at a perceptual level. Like the diminished cognitive load seen in physicians as they progress along the expertise continuum, filtering out irrelevant visual stimuli may allow for a certain *efficiency of thought* that may also free up working memory resources for physicians in clinical environments. This *efficiency of thought* is hypothesized to be the cognitive parallel of *efficiency of motion* that characterizes expert technical skills (Wanzel et al., 2003). This area represents a promising avenue for future research. By using a mixed methods approach, future studies could utilize eye-tracking technology to quantify the proportion of time spent on task relevant and task redundant stimuli by domain experts and novices. Subsequently, cued-retrospective-interviewing techniques augmented by viewing eye-tracking videos could be used to delve into how individuals with various levels of expertise compartmentalize visual information at a perceptual level. Related tasks in domains outside individuals' areas of expertise could serve as controls.

Capitalizing on eye-tracking augmented cognitive task analysis techniques, Chapters 5 and 6 investigated the cognitive processes of both experts in real trauma resuscitations and emergency medicine learners of various experience levels in simulation-based examinations. These studies showed firsthand what clinicians (and trainees) think when working or being examined. In so doing, we have begun to move away from borrowed principles of cognitive process development from other fields into the domain of resuscitation medicine specifically. Medical educators can now feel more confident knowing what the "right" cognitive processes are in this field as well as how they might evolve in their learners with experience. Future studies could focus on understanding the cognitive processes of learners struggling to succeed in resuscitation medicine. Using an approach grounded in cognitive task analysis, it would be interesting to know whether the root of these individuals' struggle

is related to the way their cognitive processes are structured and/or to task execution. Understanding this better could have the potential to open up new avenues for targeted remediation.

Assessment in medicine can be more than performance assessment

Traditional assessment in medical education has focused on testing the knowledge, attitude and skills of medical trainees (Wass, Van der Vleuten, Shatzer, & Jones, 2001). Using methods like written examinations, direct observation, multisource assessments, OSCEs, and more recently simulation, medical educators have developed a number of ways to measure performance (Epstein, 2007). What has remained elusive, however, is understanding what lays beneath the surface in the mind of the learner, at a cognitive level.

Arguably, this *invisible* aspect of performance is crucial in resuscitation medicine. A particularly germane example has to do with anticipating pitfalls, as described in Chapter 5 of this dissertation. If a physician's plan to intubate a patient with respiratory failure moves forward without complication, it's impossible for an onlooker (or an assessor) to determine whether that physician had thought of contingencies in the event that his/her plan to intubate had failed. This contingency planning is a key marker of expert performance in resuscitation medicine that generally doesn't become evident, except in the rare circumstance of plan failure or with a deep dive into the physician's mind. The results from this dissertation suggest that the latter is possible. Though not yet actualized outside of the research realm, quantitative data about cognitive load and qualitative data about cognitive processes could be used to add to the assessment picture of a medical learner. Although not common practice in expertise research in medicine, utilizing a mixed methods approach with both quantitative and qualitative input (as was done in this dissertation) has significant advantages as this technique avoids the biases inherent to single method studies and provides a richer understanding of the phenomena at hand (Gegenfurtner, Siewiorek, Lehtinen, & Säljö, 2013).

A logical next step that derives from the work presented here, related to assessment and education, is to determine whether case complexity can be modulated in real-time based on measured learner cognitive load. Given its importance in resuscitation medicine teaching,

simulation is one particular area where this could be investigated. If the complexity of a simulated scenario could be dynamically modulated to learner cognitive load (creating simulation cases that provide enough intrinsic cognitive load to be stimulating and challenging, but not too much cognitive load where learners would exceed working memory capacity) then we may be able to have a significant impact on medical education. Our research group at Queen's University is proposing to investigate this in more detail. Capitalizing on the Queen's University "Augmented Human Performance" research cluster, we will utilize our expertise in medicine, engineering, computing, psychology and human performance to design and develop an intelligent, dynamically adaptive virtual reality/artificial reality simulation environment. Utilizing continuous monitoring of cognitive, affective, behavioural and motoric performance in conjunction with point-in-time cognitive state and readiness, the proposed platform would dynamically modulate the complexity of the environment (e.g. fidelity, scenario and task complexity, temporal compression) to optimize cognitive performance and learning outcomes.

Building on suggestions by Jarodzka, Scheiter, Gerjets, and van Gog (2010), future research could also utilize experts' visual scan paths and descriptions of their associated cognitive processes as visual scaffolding for educational purposes, especially for advanced learners. Some of these techniques have been used in the visual specialties with success. In radiology teaching, novices were found to improve their diagnostic accuracy as well as visual fixations on task-relevant areas after having watched a video showing how an expert radiologist typically interprets a CT scan using eye-tracking technology (Seppänen & Gegenfurtner, 2012). Whether this strategy is effective in the complex and dynamic setting of resuscitation medicine teaching is yet to be elucidated.

The implications of these new theoretical contributions are especially important in light of the international transition to competency-based medical education. In this new framework of teaching and assessment, competency achievement has become not only theoretically important, but the need to observe and document competency by criterion-based means grounded in a developmental perspective has become central to the progression of medical learners in their training trajectory (Holmboe et al., 2010). In order to develop these criterion-based assessment strategies in resuscitation medicine, where many cognitive processes aren't readily observable to an outsider, we require a deeper approach than that

provided by traditional assessment modalities. Understanding what's *under the hood* when physicians make decisions caring for acutely unwell patients during medical crises should give medical educators a more robust understanding of where their trainees lie along the novice-expert continuum in resuscitation cases.

Limitations

A common limitation of many of the empirical studies presented in this dissertation is the reliance on "expert-novice" comparisons. A major limitation of any expert-novice study – or any known-groups comparison study for that matter – is the issue of confounding. When assessing the validity of an instrument, differences between known groups are necessary, but not sufficient to confirm validity on their own (Cook, 2015). This being said, when the results of multiple studies, looking at numerous variables all triangulate toward established expertise theory, it's likely that the results are valid.

Related to this is the issue of the cross-sectional nature of the experimental groups. This is a limitation of many studies in expertise science and is rooted in practicality as most research studies are unable to follow learners longitudinally over a protracted period of time as they develop expertise. That being said, another study examining clinical problem solving by medical students found similar results in cross-sectional and longitudinal analyses (Neufeld, Norman, Feightner, & Barrows, 1981). Still, this unavoidable limitation weakens the definitiveness that can be ascribed to the conclusions made herein. Further limiting generalizability is the issue of data collection in only one institution. Future studies could consider repeating the methodology used here but following groups of individuals over time and comparing the results across different institutions and jurisdictions.

Another limitation of this dissertation is that we focused our analysis on the knowledge and crisis resource management skills required of physicians who practice resuscitation medicine, without analyzing their technical skills. The reason for this is twofold. Firstly, in the academic environment where the studies of this dissertation were conducted, medical trainees perform a majority of procedures, while resuscitation team leaders only become involved with technical procedural aspects of a skill when a trainee is struggling or has failed. This made analyzing this facet of expertise in resuscitation medicine difficult from a

practical standpoint. Future studies could investigate technical skills specifically by either studying the trainees actually performing these skills in the clinical realm and/or focusing on experts working in a community setting where there are no learners. Secondly, there is already a significant body of literature about expertise in technical skills performance (Reznick & MacRae, 2006), while the focus of this dissertation was on the cognitive skills that underlie resuscitation medicine expertise development.

General conclusion

The work presented in this dissertation stands on the shoulders of giants in expertise science and educational psychology. We have found numerous parallels between the way experts think in resuscitation medicine and the way experts think in other fields. In summary, we found that as expertise develops, physicians involved in resuscitation medicine progressively exert less cognitive load for particular tasks and their cognitive processes become richer and more expert-like. We have provided evidence that these changes in cognitive load and cognitive processes are not invisible, as previously thought. Instead, they are measurable and have the potential to help medical educators guide their students as they progress along the novice-expert continuum.

References

Cook, D. A. (2015). Much ado about differences: why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education*, 20(3), 829-834. doi:10.1007/s10459-014-9551-3

Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, 119(2), 166.e167-166.e116. doi:https://doi.org/10.1016/j.amjmed.2005.10.036

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine*, 2007(356), 387-396.

Ericsson, K. A. (2004). Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. *Academic Medicine*, 79(10), S70-S81.

Ericsson, K. A. (2018). Capturing Expert Thought with Protocol Analysis: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks. In A. M. Williams, A. Kozbelt, K. A. Ericsson, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (2 ed., pp. 192-212). Cambridge: Cambridge University Press.

Gegenfurtner, A., Kok, E., Geel, K., Bruin, A., Jarodzka, H., Szulewski, A., & Merriënboer, J. J. (2017). The challenges of studying visual expertise in medical image diagnosis. *Medical Education*, 51(1), 97-104.

Gegenfurtner, A., Siewiorek, A., Lehtinen, E., & Säljö, R. (2013). Assessing the quality of expertise differences in the comprehension of medical visualizations. *Vocations and Learning,* 6(1), 37-54.

Haider, H., & Frensch, P. A. (1996). The Role of Information Reduction in Skill Acquisition. *Cognitive Psychology*, 30(3), 304-337. doi:https://doi.org/10.1006/cogp.1996.0009

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 172.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., Frank, J. R., & Collaborators, I. C. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, 32(8), 676-682.

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, 20(2), 146-154. doi:https://doi.org/10.1016/j.learninstruc.2009.02.019

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid Decision Making on the Fire Ground. Proceedings of the Human Factors Society Annual Meeting, 30(6), 576-580. doi:10.1177/154193128603000616

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education,* 51(1), 114-122.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science*, 7(1), 18-27.

Neufeld, V., Norman, G., Feightner, J., & Barrows, H. (1981). Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. *Medical Education*, 15(5), 315-322.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of EducationalPpsychology*, 84(4), 429.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63-71.

Reznick, R. K., & MacRae, H. (2006). Teaching Surgical Skills — Changes in the Wind. *New England Journal of Medicine*, 355(25), 2664-2669. doi:10.1056/NEJMra054785

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A., & Pappas, J. (2013). Characterizing Novice-Expert Differences in Macrocognition: An Exploratory Study of Cognitive Work in the

Emergency Department. *Annals of Emergency Medicine*, 61(1), 96-109. doi:10.1016/j.annemergmed.2012.08.034
Seppänen, M., & Gegenfurtner, A. (2012). Seeing through a teacher's eyes improves students' imaging interpretation. *Medical Education*, 46(11), 1113-1114.
Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3), 251-296. doi:10.1023/A:1022193728205
Szulewski, A., Braund, H., Egan, R., Hall, A. K., Dagnone, J. D., Gegenfurtner, A., & van Merrienboer, J. J. (2018). Through the Learner's Lens: Eye-Tracking Augmented Debriefing in Medical Simulation. *Journal of Graduate Medical Education*, 10(3), 340-341.
Szulewski, A., Brindley, P., & Van Merrienboer, J. J. (2017). Decision-making during medical crises. In P. Brindley & P. Cardinal (Eds.), *Crisis Resource Management in Acute Care Medicine* (1st ed., pp. 36-43). Ottawa: Royal College of Physicians and Surgeons Canada.
Szulewski, A., Kelton, D., & Howes, D. (2017). Pupillometry as a Tool to Study Expertise in Medicine. *Frontline Learning Research*, 5(3), 53-63.
Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(2), 147-177.
Wanzel, K. R., Hamstra, S. J., Caminiti, M. F., Anastakis, D. J., Grober, E. D., & Reznick, R. K. (2003). Visual-spatial ability correlates with efficiency of hand motion and successful surgical performance. *Surgery*, 134(5), 750-757.
Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet*, 357(9260), 945-949. doi:https://doi.org/10.1016/S0140-6736(00)04221-5

Chapter 8

English Summary

The science of expertise has been widely discussed in the literature for decades across numerous domains. With the recent international focus toward competency-based medical education (CBME), educational scientists have been increasingly interested in expertise development in health sciences and medicine specifically. Because of its relevance across medical disciplines and its high-stakes nature, resuscitation medicine represents a particularly germane example of a medical field where more study is needed to better understand the progression of learners along the novice-expert continuum to inform teaching and assessment in a CBME context.

Chapter 1 introduced the discussion about the novice-expert continuum in resuscitation medicine and structured it around an analysis of physician cognitive load as well as physician cognitive processes. More specifically, the chapter introduced the overarching research question of this dissertation: *what changes in cognitive load and processing occur in physicians as they develop expertise within the domain of resuscitation medicine? How can these changes be measured as physicians progress along the expertise continuum?* The introductory chapter concluded with the introduction of five study-specific research questions which were answered in Chapters 2 – 6.

Chapter 2 sought to answer the following research question: *what is the relationship of cognitive load, as measured by pupillometry, to level of experience in the context of a traditional knowledge-based resuscitation medicine examination?* It has been shown that dynamic changes in pupil size, also called task-evoked pupillary responses (TEPRs), are associated with changes in cognitive processing demands. The magnitude of this change is thought to be a reliable marker of cognitive load. In this chapter, we presented the first empirical study of this dissertation that investigated the use of pupillometry in cognitive load measurement in a medical testing environment. In this study, twenty emergency medicine trainees were divided into novice and trained physician groups. We assessed changes in pupil diameter as participants answered arithmetic questions, general knowledge questions, and clinical emergency medicine questions in a controlled setting.

Difficult arithmetic questions caused greater changes in TEPRs than easy ones. TEPRs were similar between groups when answering general knowledge questions but were significantly greater for novices than trained physicians when answering clinical questions. TEPRs in trained physicians were significantly greater when answering difficult clinical questions than

easy ones, whereas TEPRs in novices were similar. For those clinical questions answered correctly by both groups, TEPRs in novices were greater than those in trained physicians despite all participants answering correctly.

The results of the data presented in Chapter 2 suggest that novices experienced greater cognitive load to answer clinical questions than trained physicians, even when both responded correctly. We concluded that measuring TEPRs has the potential to be a valuable assessment tool in the context of traditional testing in resuscitation medicine by providing novel, objective measures of cognitive load.

Chapter 3 aimed to answer the following research question: *in the context of a resuscitation medicine examination, what is the validity of using a physiologic measure of cognitive load (pupillometry) and a psychometric one (survey item) as markers of cognitive load among physicians with different levels of experience?* The approach to the quantification of cognitive load has focused primarily on physiologic and psychometric measures. Although the construct measured by both metrics is thought to represent overall cognitive load, there is a paucity of studies that compares these techniques head-to-head. Grounded in a contemporary validity framework, in this chapter we compared data obtained from one physiologic tool (pupillometry) to one psychometric tool (Paas scale) to explore whether the tools actually measured the construct of cognitive load as purported. Thirty-two participants with a range of resuscitation medicine experience completed resuscitation-medicine based multiple-choice-questions as well as arithmetic questions. As expected, cognitive load (as measured by both tools) was found to be higher for the more difficult questions as well as for questions that were answered incorrectly. The group with the least medical experience had higher cognitive load than both the intermediate and experienced groups when answering domain-specific questions. There was a strong positive correlation between the two cognitive load measurement tools. These findings support the validity argument that both physiologic and psychometric metrics measure the construct of cognitive load in a resuscitation-medicine testing environment.

The research question answered in Chapter 4 was: *in resuscitation-based simulation OSCE's, how (and to what extent) are particular gaze patterns of residents associated with exam performance? How do these gaze patterns vary across scenarios?* Information gathering is a key responsibility of the physician team leader managing a resuscitation case. With this in

mind, this chapter moved away from cognitive load measurement and into the realm of studying physician visual behaviours. In an effort to study information gathering propensities of physicians, emergency medicine residents were outfitted with gaze-tracking glasses during two simulation-based examinations (29 and 13 residents respectively). Blinded experts assessed video-recorded performances using a simulation performance assessment tool. The relationships between gaze patterns and performance scores were analyzed.

Specific gaze patterns were found to be correlated with objective performance. In line with the information-reduction hypothesis, high performers were found to more often fixate on task-relevant stimuli and appropriately ignore task-irrelevant stimuli compared with lower performers.

The results presented in this chapter suggest that there may be objective ways to quantify visual expertise in the context of resuscitation medicine. These findings may allow for better characterization of expertise development in resuscitation medicine and may provide the foundation for future study of physician visual behaviours in simulated and real resuscitation cases.

The study reported in Chapter 5 aimed to answer the following question: *what are the cognitive processes of expert physicians while leading actual trauma resuscitations?* Crisis resource management skills are integral to leading the resuscitation of a critically ill patient. Despite their importance, crisis resource management skills (and their associated cognitive processes) have traditionally been difficult to study in the real world. Exploring these cognitive processes was the focus of this chapter. In this chapter, we derived key cognitive processes underpinning expert performance in resuscitation medicine using a new eye-tracking based video capture method during clinical cases.

A total of 10 trauma resuscitations led by 4 expert trauma team leaders was analyzed. The physician team leaders were outfitted with mobile eye-tracking glasses for each case. After each case, participants were debriefed using a modified cognitive task analysis augmented by viewing their own first-person-perspective eye-tracking video from the clinical encounter.

Five major themes were derived from the interviews: logistic awareness, managing uncertainty, visual fixation behaviors, selective attendance to information, and anticipatory behaviors. We concluded that a better understanding of these cognitive processes has the potential to enhance educational methods and to create new assessment modalities of these previously tacit aspects of expertise in this field.

The following research question was answered in Chapter 6: *what are the cognitive processes of medical trainees in simulation-based resuscitation examinations? How do these cognitive processes vary with examination performance?* Resuscitation medicine education has relied heavily on simulation-based training to aid in learners' cognitive development and approach to managing patients who are acutely unwell. Initially used as a teaching method, simulation has recently been gaining a greater role in trainee assessment. Although some elements of performance in a medical simulation are readily visible to an outside observer and are measurable with traditional tools, the cognitive processes underlying crisis resource management are not. As a result, our understanding of these cognitive processes (including how they may differ over the continuum of learner performance) is incompletely understood. Chapter 6 of this dissertation describes an empirical study in which twenty-two emergency medicine residents completed a resuscitation-based examination in the simulation laboratory while wearing eye-tracking glasses. Resident performance was assessed by a blinded expert using an entrustment-based scoring tool. Following the examinations, residents were interviewed using a retrospective debrief protocol cued by viewing their own performance using the first-person video generated by the eye-tracker. Higher performing residents were better able to anticipate, selectively attend to relevant information, manage cognitive demands, and took a concurrent (as opposed to linear) approach to managing the simulated patient. The results provide insight into the evolution of the cognitive processes of residents as they become more proficient in resuscitating simulated patients in an examination environment.

The final chapter of this dissertation summarized the main findings of each of the empirical studies presented in Chapters 2 – 6 and related these findings back to the study-specific research questions as well as to the overarching research question of the dissertation. The chapter then delved into some new theoretical contributions.

First, the phenomenon of understanding expertise development in resuscitation medicine was discussed. Moving beyond traditional notions of performance, this section focused on *getting under the hood* of what it means to be an expert in this field. The reported insights of the various analyses of cognitive load and cognitive processes within this dissertation open new avenues for research and scientific enquiry. In addition, medical educators can now feel more confident knowing how cognitive load changes with the experience of their learners as well as what the underlying cognitive processes are in resuscitation medicine and how they might evolve with experience.

Next, the idea that assessment in medicine can be more than simply performance assessment was introduced. Armed with a new understanding of some of the previously tacit elements of resuscitation medicine practice, medical educators may now be better able to accurately document learner progression along the novice-expert continuum in this field by expanding on traditional assessment metrics with some of the novel data on cognitive load and cognitive processes that this dissertation has introduced. This is particularly relevant in a specialty like resuscitation medicine where many of the cognitive processes that underlie expertise in this domain are not generally visible (and therefore able to be assessed) by an outside observer.

The general discussion concluded with a discussion about limitations, which included the issues of expert-novice comparisons, the cross-sectional nature and the single-study-centre design of the experiments.

Taken together, the studies reported in this dissertation have shown that expertise development in resuscitation medicine can be reliably described, and in some instances quantified, despite the inherent limitations of a medical specialty in which many decisions are processed at a cognitive level. By capitalizing on both quantitative and qualitative methodology, this thesis has added to the richness of the conversation about our understanding of expertise development in resuscitation medicine. This is a conversation that is both needed and timely given the climate around patient safety and the shift toward competency based medical education.

Chapter 9

Nederlandse samenvatting

De wetenschap van de expertise wordt al tientallen jaren uitgebreid in vele domeinen in de literatuur besproken. Door de recente internationale focus op competentiegericht medisch onderwijs (CBME) hebben onderwijswetenschappers steeds meer belangstelling gekregen voor expertiseontwikkeling in de gezondheidswetenschappen, met name in de geneeskunde. Omdat zij op alle medische specialismen betrekking heeft en er veel bij op het spel staat, is reanimatiegeneeskunde een bijzonder relevant voorbeeld van een medisch gebied waarbinnen meer onderzoek nodig is om meer inzicht te verkrijgen in de voortgang van studenten langs het nieuweling-expertcontinuüm ten behoeve van het onderwijs en toetsing in een CBME-context.

Hoofdstuk 1 introduceerde de discussie over het nieuweling-expertcontinuüm in reanimatiegeneeskunde en structureerde deze rondom een analyse van cognitieve belasting en cognitieve processen bij artsen. Meer specifiek introduceerde het hoofdstuk de overkoepelende onderzoeksvraag van dit proefschrift: Welke veranderingen in cognitieve belasting en verwerking vinden er plaats bij artsen naarmate zij expertise ontwikkelen binnen het domein van de reanimatiegeneeskunde? Hoe kunnen deze veranderingen gemeten worden terwijl zij langs het expertisecontinuüm vorderen? Het inleidend hoofdstuk sloot af met de aankondiging van vijf studiegerichte onderzoeksvragen die in Hoofdstuk 2 t/m 6 beantwoord werden.

Hoofdstuk 2 poogde de volgende onderzoeksvraag te beantwoorden: Wat is het verband tussen (door pupillometrie gemeten) cognitieve belasting en de mate van ervaring in het kader van een traditionele kennistoets over reanimatiegeneeskunde? Het is aangetoond dat dynamische veranderingen in pupilgrootte, ook wel “door de taak opgewekte pupilresponsen” (TEPRs) genoemd, in verband worden gebracht met veranderingen in de mate waarin er een beroep gedaan wordt op ons cognitieve verwerkingsvermogen. De grootte van deze verandering wordt beschouwd als een betrouwbare indicator voor cognitieve belasting. In dit hoofdstuk presenteerden we de eerste empirische studie van dit proefschrift die het gebruik van pupillometrie bij het meten van cognitieve belasting in een geneeskundetoetsomgeving onderzocht. In deze studie werden 20 aiossen spoedeisende geneeskunde onderverdeeld in een groep “nieuwelingen” en een groep “opgeleide artsen”. We beoordeelden of er veranderingen optraden in pupildiameter terwijl de deelnemers

rekenkundige vragen, algemene kennisvragen en vragen over klinische spoedeisende geneeskunde beantwoordden in een gecontroleerde omgeving.

Moeilijke rekenkundige vragen zorgden voor grotere veranderingen in TEPRs dan makkelijke vragen. De TEPRs waren voor beide groepen gelijk bij de beantwoording van algemene kennisvragen, maar waren significant groter voor de nieuwelingen dan voor de opgeleide artsen bij de beantwoording van klinische vragen. De TEPRs waren significant groter wanneer de opgeleide artsen moeilijke klinische vragen beantwoordden dan wanneer dit makkelijke vragen betrof, terwijl de TEPRs bij de nieuwelingen gelijk bleven. Wat de klinische vragen die door beide groepen juist werden beantwoord betreft, waren de TEPRs groter bij de nieuwelingen dan bij de opgeleide artsen, ondanks het feit dat alle deelnemers de vragen juist hadden beantwoord. De resultaten van de in Hoofdstuk 2 gepresenteerde data maken aannemelijk dat de nieuwelingen meer cognitieve belasting ervoeren bij het beantwoorden van klinische vragen dan de opgeleide artsen, zelfs wanneer beide groepen het antwoord juist hadden. We concludeerden dat het meten van TEPRs een nuttig toetsinstrument kan zijn op het gebied van traditioneel toetsen in reanimatiegeneeskunde, doordat het nieuwe, objectieve metingen van cognitieve belasting verschaft.

Hoofdstuk 3 beoogde de volgende onderzoeksvraag te beantwoorden: Welke validiteit heeft het gebruik van een fysiologische maatstaf voor cognitieve belasting (pupillometrie) en een psychometrische maatstaf (enquête-item) als indicatoren voor cognitieve belasting bij artsen met verschillende ervaringsniveaus in het kader van een reanimatiegeneeskundetoets? Voor het meten van cognitieve belasting richt men zich hoofdzakelijk op fysiologische en psychometrische maatstaven. Hoewel wordt aangenomen dat het construct dat beide meeteenheden meten de totale cognitieve belasting omvat, zijn er te weinig studies die deze technieken effectief met elkaar hebben vergeleken. In dit hoofdstuk dat op een hedendaags validiteitskader berustte, vergeleken we de data die we uit een fysiologisch instrument verkregen (pupillometrie) met een psychometrisch instrument (de schaal van Paas) om te onderzoeken of de instrumenten daadwerkelijk het construct “cognitieve belasting” maten zoals werd beweerd. Tweeëndertig deelnemers met verschillende mate van ervaring op het gebied van reanimatiegeneeskunde vulden zowel meerkeuzevragen als rekenkundige vragen over reanimatiegeneeskunde in. Zoals verwacht, bleek de (door beide instrumenten gemeten) cognitieve belasting groter te zijn bij de

moeilijkere vragen, alsook bij de vragen die onjuist werden beantwoord. De groep met de minste geneeskundeervaring liet bij het beantwoorden van de domeinspecifieke vragen een grotere cognitieve belasting zien dan de halfgevorderde en ervaren groepen.

Er was een sterke positieve correlatie tussen de twee instrumenten die cognitieve belasting maten. Deze bevindingen staven het validiteitsargument dat zowel fysiologische als psychometrische meeteenheden het construct “cognitieve belasting” in een toetsomgeving binnen reanimatiegeneeskunde meten.

De onderzoeksvraag die in Hoofdstuk 4 werd beantwoord was: Hoe (en in hoeverre) houden bepaalde kijkpatronen van aiossen verband met hun toetsprestaties bij op reanimatiesimulaties beruste stationstoetsen? Hoe verschillen deze kijkpatronen van scenario tot scenario? Het verzamelen van informatie is een belangrijke verantwoordelijkheid van de arts-teamleider die een reanimatiepatiënt behandelt. Met dit in het achterhoofd nam dit hoofdstuk afstand van cognitieve-belastingsmetingen en richtte zich in plaats daarvan op het bestuderen van visuele gedragingen door artsen. In een poging om de geneigdheid van artsen om informatie te verzamelen te onderzoeken, werden aiossen spoedeisende geneeskunde tijdens twee simulatietoetsen (respectievelijk 29 en 13 aiossen) voorzien van een bril die voortdurend hun blikveld registreerde. Geblindeerde experts beoordeelden de op video vastgelegde prestaties met behulp van een simulatieprestatiebeoordelingsinstrument. De verbanden tussen kijkpatronen en prestatiescores werden geanalyseerd.

Specifieke kijkpatronen bleken te correleren met objectieve prestaties. In overeenstemming met de informatie-reductiehypothese bleken de betere presteerders hun aandacht vaker te vestigen op taakrelevante stimuli en de taakirrelevante stimuli op passende wijze te negeren ten opzichte van de minder goede presteerders.

De in dit hoofdstuk gepresenteerde resultaten maken aannemelijk dat er mogelijk objectieve manieren zijn om visuele expertise in het kader van reanimatiegeneeskunde te meten. Deze bevindingen zouden een betere beschrijving van expertiseontwikkeling in reanimatiegeneeskunde mogelijk kunnen maken en de basis kunnen vormen voor

toekomstige studies over visuele gedragingen door artsen bij gesimuleerde en echte reanimatiepatiënten.

De in Hoofdstuk 5 beschreven studie beoogde de volgende onderzoeksvraag te beantwoorden: Welke cognitieve processen maken specialisten door bij het begeleiden van reanimaties van echte traumapatiënten? Crisis-resource-managementvaardigheden vormen een wezenlijk onderdeel van de begeleiding van een reanimatie van een patiënt in kritieke toestand. Ondanks het belang ervan zijn crisis-resource-managementvaardigheden (en de bijbehorende cognitieve processen) van oudsher moeilijk te onderzoeken in de echte wereld. Het onderzoeken van deze cognitieve processen stond centraal in dit hoofdstuk. In dit hoofdstuk distilleerden we belangrijke cognitieve processen die ten grondslag liggen aan de prestaties van reanimatiegeneeskundespecialisten door gebruik te maken van een nieuwe, op eye tracking beruste methode voor het vastleggen van video tijdens klinische ziektegevallen. Er werden in totaal 10 reanimaties van traumapatiënten die onder de supervisie van vier specialist-traumateamleiders werden uitgevoerd, geanalyseerd. De arts-teamleiders werden voor elk geval voorzien van een mobiele eye-trackingbril. Na elk geval werden de deelnemers ondervraagd aan de hand van een aangepaste cognitieve taakanalyse, terwijl ze de eye-tracking video-opname van het klinische ziektegeval, gezien vanuit hun eigen perspectief, terugzagen.

Er werden vijf centrale thema's uit de interviews gedistilleerd: logistiek bewustzijn, omgaan met onzekerheid, visuele fixeergedragingen, selectieve aandacht voor informatie en anticiperende gedragingen. We concludeerden dat een beter begrip van deze cognitieve processen ons in staat stelt onderwijsmethoden te verbeteren en nieuwe methodieken te ontwikkelen waarmee deze expertiseaspecten die voorheen impliciet waren in dit gebied getoetst kunnen worden.

De volgende onderzoeksvraag werd beantwoord in Hoofdstuk 6: Welke cognitieve processen maken aiossen door tijdens op simulatie beruste reanimatietoetsen? Hoe hangen deze cognitieve processen samen met toetsprestaties? Om studenten te ondersteunen bij hun cognitieve ontwikkeling en de wijze waarop zij acuut zieke patiënten behandelen maakt het reanimatiegeneeskundeonderwijs veel gebruik van simulatieonderwijs. Simulatie, aanvankelijk gebruikt als onderwijsmethode, heeft onlangs een grotere rol gekregen bij de

toetsing van aiossen. Hoewel bij een medische simulatie sommige prestatieaspecten duidelijk zichtbaar zijn voor een externe observator en deze met traditionele methoden gemeten kunnen worden, geldt dit niet voor de cognitieve processen die aan crisis resource management ten grondslag liggen. Dientengevolge is ons begrip van deze cognitieve processen (en ook hoe zij over het continuüm van studentprestaties kunnen verschillen) onvolledig. Hoofdstuk 6 van dit proefschrift beschrijft een empirische studie waarin tweeëntwintig aiossen spoedeisende geneeskunde in het simulatielaboratorium een reanimatietoets maakten terwijl zij een eye-trackingbril droegen. De prestaties van aiossen werden door een geblindeerde expert beoordeeld aan de hand van een entrustment-scorelijst. Na afloop van de toets interviewden we de aiossen met behulp van een retrospectief interviewprotocol. Als geheugensteuntje kregen de aiossen daarbij hun eigen prestaties op het door de eye tracker gemaakte videofilmpje vanuit hun eigen perspectief terug te zien. De beter presterende aiossen waren beter in staat te anticiperen, hun aandacht selectief te vestigen op relevante informatie, om te gaan met cognitieve eisen en een gelijktijdige (in plaats van een lineaire) aanpak te hanteren ten aanzien van de behandeling van de simulatiepatiënt. De resultaten bieden inzicht in de ontwikkeling van cognitieve processen bij aiossen naarmate zij steeds vaardiger worden in het reanimeren van simulatiepatiënten in een toetsomgeving.

Het laatste hoofdstuk van dit proefschrift vatte de belangrijkste bevindingen van elk van de in Hoofdstuk 2 t/m 6 gepresenteerde empirische studies samen en koppelde deze bevindingen terug aan zowel de studiegerichte onderzoeksvragen als de overkoepelende onderzoeksvraag van het proefschrift. Het hoofdstuk verdiepte zich vervolgens in enkele nieuwe theoretische bijdragen.

Eerst werd het terrein van “hoe expertiseontwikkeling in reanimatiegeneeskunde te begrijpen”, besproken. Deze paragraaf ging verder dan de traditionele begrippen van prestaties en stond uitgebreid stil bij de vraag “wat betekent het om een expert te zijn op dit gebied?”. De inzichten in de verschillende analyses van cognitieve belasting en cognitieve processen die in dit proefschrift werden gerapporteerd, bieden nieuwe ingangen voor onderzoek en wetenschappelijke vraagstellingen. Bovendien kunnen medische opleiders zich nu zekerder voelen doordat zij weten hoe de cognitieve belasting van hun studenten verandert naargelang zij ervaring opdoen, alsook welke achterliggende cognitieve

processen er spelen bij reanimatiegeneeskunde en hoe deze zich mogelijk ontwikkelen naarmate hun studenten ervaring opdoen.

Vervolgens werd het idee aangeroerd dat toetsing bij geneeskunde meer kan behelzen dan alleen maar toetsing van prestaties. Gewapend met een nieuw begrip van enkele, voorheen impliciete, aspecten binnen de reanimatiegeneeskundepraktijk, zullen medische opleiders nu wellicht beter in staat zijn om nauwkeurig de voortgang van hun studenten langs het nieuweling-expertcontinuüm op dit gebied te bepalen. Dit door de traditionele meeteenheden voor toetsing uit te breiden met enkele van de door dit proefschrift geïntroduceerde nieuwe gegevens omtrent cognitieve belasting en cognitieve processen. Dit is met name van belang bij een specialisme zoals reanimatiegeneeskunde, waarbij veel van de cognitieve processen die ten grondslag liggen aan expertise op dit gebied doorgaans niet zichtbaar zijn voor een externe observator (en daardoor ook niet door hem/haar kunnen worden beoordeeld).

De algemene discussie sloot af met een bespreking van de beperkingen, waaronder die met betrekking tot expert-nieuwelingvergelijkingen, de transversale aard en het één-instellingsontwerp van de experimenten.

Alles samengenomen hebben de in dit proefschrift uiteengezette studies aangetoond dat expertiseontwikkeling bij reanimatiegeneeskunde op betrouwbare wijze beschreven en in sommige gevallen gemeten kan worden, ondanks de beperkingen inherent aan een medisch specialisme waarbinnen veel besluiten op een cognitief niveau worden verwerkt. Door gebruik te maken van zowel kwantitatieve als kwalitatieve methodieken, heeft dit proefschrift voor een verrijking gezorgd van de discussie over ons begrip van expertiseontwikkeling in reanimatiegeneeskunde. Deze discussie is niet alleen noodzakelijk, maar komt ook op het juiste moment, gezien het klimaat rondom patiëntveiligheid en de omschakeling naar competentiegericht medisch onderwijs.

Valorization

Common to any academic pursuit is the generation of knowledge to expand our understanding of science. In addition to this contribution, the implications of this dissertation have the added potential for real-world application in the short-to-medium-term.

Relevance

With varying frequency, physicians across the spectrum of almost all medical specialties are involved, to some degree, in resuscitating their patients. There is now a public expectation that physicians trained in all residency programs will demonstrate basic competence in this area. With the recent move toward competency-based medical education (CBME), an increasing number of residency programs have begun formalizing this expectation in teaching and assessment of their medical trainees.

Though basic competence is necessary, the goal of any residency program should be on supporting trainees as they progress toward expertise along the novice-expert continuum. Prior to undertaking this PhD thesis, we did not have a good understanding of how expertise in resuscitation medicine developed, nor were we able to reliably measure it. A key reason for this is that many of the decisions made by a physician while resuscitating a patient in a high-stakes and time-constrained environment occur at a cognitive level. As a result, they have traditionally been difficult to measure and describe.

The studies presented in this dissertation describe methods of getting beyond this challenge. Grounded in analyses of physician cognitive load and cognitive processes, we have shown that progression along the novice-expert continuum in resuscitation medicine can be described and measured reliably.

These findings are timely and relevant given the move toward CBME and in the current era of patient safety.

Target groups, activities and products

The results of the studies presented in this dissertation will be of interest to a number of groups. In particular, the results will be of interest to medical educators. Up until now, teaching and assessing the crisis resource management (CRM) skills that underpin physician performance during a resuscitation case has been mostly *ad hoc*. Simulation has provided a vehicle to start teaching and assessing these skills, but most simulation training has relied on exposing medical learners to a high volume of cases and assuming they will learn CRM skills with this exposure. What has been lacking is a complete understanding of how these skills are naturally acquired by learners as they progress along the novice-expert continuum in resuscitation medicine. As a result, much of the teaching and assessment practices in resuscitation medicine have taken a one-size-fits-all approach. By understanding how physicians' cognition changes as they gain expertise, medical educators should now have a more nuanced approach to teaching and assessing their learners.

Moreover, armed with new knowledge about how expertise develops in resuscitation medicine, medical educators may take the insights afforded by the studies presented here to design thoughtful competence-based curricula. Knowing how experts make decisions during medical crises will allow educators to focus on these elements specifically during simulation design and debriefing. In addition, this knowledge may allow for the creation of more accurate learning objectives and assessment frameworks in a CBME context. This, in turn, will lead to a more accurate characterization of residents' abilities in resuscitation medicine.

Resident learners themselves will also find the results of this dissertation informative. We have described, in detail, how cognitive load and cognitive processes change as a resident's knowledge matures in resuscitation medicine. We have also provided an overview of these processes in experts themselves. With an appreciation for this cognitive maturation process, physician learners should be better able to assess where they currently lay on the novice-expert continuum. This knowledge can help learners to set appropriate, realistic and relevant learning goals as they will now better understand what they are specifically striving for. This type of self-reflection activity has the potential to motivate residents and ensure they remain engaged in learning over the long-term.

Finally, corporate entities may have an interest in quantifying expertise in medicine. As we move toward the next generation of simulation, there is an appetite for increased automation and a more tailored learning experience for individual trainees. At the same time, financial constraints of medical schools and time constraints of medical educators are becoming important factors. One key challenge in creating simulations that are adaptive to a learner's ability is being able to measure that ability accurately. Building on the deeper understanding of expertise in resuscitation medicine that we have presented in this dissertation, corporations will now have better tools to start to solve these problems. If simulation can be built to respond (using artificial intelligence) to an individual learner's measured level of expertise, then there is a potential to create learning that is more targeted and that relies to a lesser degree on the limited time of medical educators. This type of technology will be needed by educational institutions, and as a result, will be attractive from a corporate investment perspective.

Innovation

We have shown that expertise in resuscitation medicine can be reliably measured and described using a variety of methodologies, even though many decisions in this field are made and processed at a cognitive level. One novel example of this type of methodology, presented in this dissertation, is the introduction of eye-tracking devices to measure expertise development and to support training and assessment in medicine.

These innovative methodologies and the subsequent insights that they afforded were made possible by bringing together knowledge from the domains of medicine, educational psychology and expertise science. Presenting this work at conferences and in peer-reviewed publications from each of these respective domains has contributed to reminding educators of the importance of grounding their practices in scientific theory, while at the same time ensuring that science continues to answer the questions that are most relevant to the needs of medical educators. The marriage of these interests is primarily what makes the results of this dissertation innovative.

Schedule and implementation

A number of initiatives have borne out of the results of the studies of this dissertation. For example, based on the finding that physicians develop an ability to appropriately de-prioritize irrelevant information while prioritizing what's relevant during a resuscitation case has led to the implementation of realistic distractors in simulation-based resuscitation examinations at our institution. This has allowed educators to more accurately assess residents and has opened up avenues for new discussion during simulation debriefs. Anecdotally, residents have mentioned that this has been beneficial for their learning.

In addition, we have started to use eye-tracking augmented debriefing, as was introduced in this dissertation, into some of our local simulation activities. This has been well-received by our learners who find that the technique adds to their educational experience.

Finally, we have created a research collaborative, called the Centre for Augmented Human Performance at Queen's University. This group brings together experts in medicine, simulation, engineering, education and social sciences as it strives to be the leader of the next generation of simulation training. This group has partnered with virtual reality, augmented reality and educational corporations on a number of projects as well as grant applications. The work that has come out of this dissertation serves as the starting point for many of the projects that this collaborative is currently working on.

The cost of these initiatives is relatively high, especially in the short-term. However, as the technology advances and requires less time of medical educators, the expectation is that the overall cost savings of this program will be substantive.

Acknowledgements

Much like the practice of resuscitation medicine itself, completing a PhD is a team effort.

First and foremost, thank you to my wife, Libby. You are my rock. You thought that doing this PhD was a good idea from the beginning and you've made it possible throughout the process by motivating me and being the glue of our family. To Joseph and Jack (and to our new baby who will be born around the time I'm defending this dissertation) – you have been what keeps me grounded and reminds me what's important in life.

Thank you, Jeroen. The opportunity to work with you has had a profound impact on my career. When I started my MHPE, I had an interest in cognitive load theory, but I would never have expected this interest to evolve into a PhD and a career focused on research. Your insights during our many discussions over the years have always been thoughtful, relevant and impactful. Thank you for seeing the potential and motivating me to dive into this PhD.

Thank you, Andreas. Your insights and ideas can be found throughout each part of this dissertation. You have challenged me to look outside my own area of study and, as a result, you have helped me connect ideas from theoretical domains I would never have considered. Also, I appreciate you creating opportunities for me to get involved in some of your own work.

Thank you to the Faculty in the Department of Emergency Medicine at Queen's University for seeing the value in this work and for your ongoing support of this research program. It's greatly appreciated. And thank you to the Faculty of Health Sciences at Queen's University who supported this work through a number of grants.

Finally, a special thank you to my mentor, Dan. Before we met on a RACE call in the middle of the night in 2011 – that ultimately led to a case report we published together – I had no interest in research. Over the subsequent years, you gave me opportunity after opportunity to become involved in your work. And I became hooked. You have single-handedly changed the direction of my career. Thank you for your mentorship – both professional and personal.

SHE Dissertations Series

The SHE Dissertation Series publishes dissertations of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD theses at Maastricht University. The most recent ones are listed below. For more information go to:

https://she.mumc.maastrichtuniversity.nl

Amalba, A. (20-12-2018) Influences of problem-based learning combined with community-based education and service as an integral part of the undergraduate curriculum on specialty and rural workplace choices

Melo, B. (12-12-2018) Simulation Design Matters; Improving Obstetrics Training Outcomes

Olmos-Vega, F. (07-12-2018) Workplace Learning through Interaction: using socio-cultural theory to study residency training

Chew, K. (06-12-2018) Evaluation of a metacognitive mnemonic to mitigate cognitive errors

Sukhera, J. (29-11-2018) Bias in the Mirror. Exploring Implicit Bias in Health Professions Education

Mogre, V. (07-11-2018) Nutrition care and its education: medical students' and doctors' perspectives

Ramani, S. (31-10-2018) Swinging the pendulum from recipes to relationships: enhancing impact of feedback through transformation of institutional culture

Winslade N. (23-10-2018) Community Pharmacists' quality-of-care metrics. A prescription for improvement

Eppich, W. (10-10-2018) Learning through Talk: The Role of Discourse in Medical Education

Wenrich, M. (12-09-2018) Guided Bedside Teaching for Early Learners: Benefits and Impact for Students and Clinical Teachers

Marei, H. (07-09-2018) Application of Virtual Patients in Undergraduate Dental Education

Waterval, D. (26-04-2018) Copy but not paste, an exploration of crossborder medical curriculum partnerships

Smirnova, A. (04-04-2018) Unpacking quality in residency training and health care delivery

Hikspoor, J. (05-12-2017) Development of the heart and vessels in the caudal part of the human body

Boymans, T. (06-10-2017) Hip arthroplasty in the very elderly: anatomical and clinical considerations

Zaidi, Z. (04-10-2017) Cultural hegemony in medical education: exploring the visibility of culture in health professions

Harrison, C. (20-09-2017) Feedback in the context of high-stakes assessment: can summative be formative?

Mekonen, H. (30-06-2017) Development of the axial musculo-skeletal system in humans

Taylor, T. (29-03-2017) Exploring Fatigue as a Social Construct: Implications for Work Hour Reform in Postgraduate Medical Education

McLellan, L. (29-03-2017) Prescribing the right medicine: Perspectives on education and practice

Ignacio, J. (09-02-2017) Stress Management in Crisis Event Simulations for Enhancing Performance

Bolink, S. (19-01-2017) Functional outcome assessment following total hip and knee arthroplasty; Implementing wearable motion sensors

Appendix A

Pupillometry as a tool to study expertise in medicine

Published as: Szulewski, A., Kelton, D., & Howes, D. (2017). Pupillometry as a Tool to Study Expertise in Medicine. *Frontline Learning Research*, 5(3), 53-63.

Abstract

Background
Pupillometry has been studied as a physiological marker for quantifying cognitive load since the early 1960s. It has been established that small changes in pupillary size can provide an index of the cognitive load of a participant as he/she performs a mental task. The utility of pupillometry as a measure of expertise is less well established, although recent research in the fields of education, medicine and psychology indicates that differences in pupillary size during domain-specific tasks allows differentiation between experts and novices in appropriately designed experiments.

Purpose
The goal of this review is to explore the existing body of evidence for the use of pupillometry as a measure of expertise and to identify its strengths and constraints within the context of expertise research in the medical sciences.

Results
Pupillometry is a robust metric that allows researchers to better understand cognitive load in medical practitioners with varying levels of expertise. In medical expertise research, it has been used to study surgeons, anesthetists and emergency physicians. Its strengths include its ability to provide quantitative and objective outputs, to be measured unobtrusively with new technology and to be precisely computed as cognitive load changes over the course of completion of a task. Constraints associated with this methodology include its potential inaccuracy with changes in ambient light and pupillary accommodation as well as the need for relatively expensive equipment.

Conclusion
With recent technological advances, pupillometry has become a simple and robust method for quantifying physiological changes attributable to cognitive load and is increasingly being utilized in medical education. It can be used as a reliable marker of mental effort and has been shown to differentiate levels of expertise in medical practitioners.

Background

The measurement of human cognitive load has been of interest to researchers for decades. Knowing *how intensely* a person is thinking has implications beyond knowing *what* that person is thinking about. This is particularly relevant in the context of professional domains (like medicine) where critical and cognitively loading decisions often need to be made with limited time and in the context of other competing priorities.

This "intensity" of thinking, which is related to cognitive load, functions within the constraints of a limited working memory. Working memory is a key executive function. Executive functions are a group of mental processes that are required when an individual has to pay attention, and when it would be considered inappropriate, insufficient or impossible to rely on instinct or to respond automatically (Burgess & Simons, 2005). In addition to working memory, executive functioning involves two other core activities: inhibition (self-control, selective attention, cognitive inhibition) and cognitive flexibility (which is closely related to creativity). These three core activities are combined in different ways to build higher order executive functions such as reasoning, problem solving and planning (Diamond, 2013).

Working memory, which is responsible for the manipulation of stored information (or our ability to "think"), is generally thought to be limited (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). Ongoing work in this area now suggests that with experience, experts are able to expand working memory capacity by having developed methods for storage and retrieval of domain-specific information in long-term memory – so called long-term working memory (Ericsson & Kintsch, 1995). This is accomplished, in part, by pattern recognition and schema development, and a resultant relative decrease in the cognitive load that a problem or situation imposes as an individual becomes more expert-like (Szulewski, Roth, & Howes, 2015).

Functionally, cognitive load can be thought of as the mental capacity that is allocated to performing a task (Paas et al., 2003). It is thought to be comprised of three components: intrinsic cognitive load (ICL), extraneous cognitive load (ECL) and germane cognitive load (GCL) (Young, Van Merrienboer, Durning, & Ten Cate, 2014). ICL is a function of expertise and task complexity, while ECL is related to suboptimal information presentation conditions.

GCL refers to the working memory resources that are dedicated to processing ICL, and thus to learning (Sweller, 2010).

In general, researchers measure cognitive load using psychometric scales, physiological variables and secondary task methodology (Paas et al., 2003). Briefly, psychometric scales gather subjective data from participant self-reports after task completion. Physiological variables use task-evoked pupillary responses (TEPRs) or pupillometry, heart rate variability, galvanic skin response (among others) as surrogate markers of cognitive load. Secondary task methodology relies on participants' performance on a secondary task (that requires sustained attention, like detecting an auditory signal) and uses this information to glean the level of cognitive load imposed by the primary task. Each of these techniques has its own strengths and limitations; but in general, each is thought to provide data about total (or measurable) cognitive load. The contribution of intrinsic, extraneous and germane cognitive load to each of these measurement techniques remains to be elucidated (Leppink, Paas, van Gog, van der Vleuten, & van Merriënboer, 2014).

This review focuses on one particular physiological method of measuring cognitive load – pupillometry, which is the study of changes in pupil size. We will first examine the technique of pupillometry as a surrogate marker for cognitive load in non-medical domains and then we will focus the discussion on pupillometry research in medicine and how this relates to the development of expertise. Finally, constraints of the technique will be discussed.

Pupil physiology

Dilation and constriction of the human pupil is necessary for day-to-day visual tasks. Dilation (mydriasis) of the pupil is accomplished by the contraction of the iris dilator (radial) muscle, which is controlled by the sympathetic nervous system. Constriction (miosis) of the pupil occurs when the iris sphincter (circular) muscle contracts, which is controlled by the parasympathetic nervous system. Two commonly tested reflexes in clinical medicine are the light reflex and the accommodation reflex. During the light reflex, the pupil dilates in low luminance environments and constricts in high luminance environments. In the

accommodation reflex, as an individual changes visual focus from a distant object to a closer object, the pupil constricts, and vice-versa (Lang, 2015).

In addition to these clinically measurable and commonly discussed reflexes, pupils also change in size as a result of non-visual stimuli. This was first described in detail by Hess and Polt (1960) where it was shown that pupil size varied when participants viewed particular images (for example, sexually suggestive ones). Follow-up studies by this group and others further demonstrated that pupil size could be used to measure cognitive load (or mental effort). Physiologically, it is thought that pupil size changes with cognitive loading as a result of pathways that originate in the locus coeruleus, which is a major norepinephrine source in the brain (Laeng, Sirois, & Gredebäck, 2012). In fact, locus coeruleus activity has been shown to be very closely related to sympathetic activity and changes in pupil size (Aston-Jones & Cohen, 2005). These pupillary responses are spontaneous and very difficult to control voluntarily. The voluntary dilation of a subject's pupils is only possible indirectly if the subject imagines a situation (e.g. self-induced sexual imagery) where his/her pupils would normally dilate (Whipple, Ogden, & Komisaruk, 1992). This would be particularly difficult, if not impossible, to systematically do while simultaneously performing other cognitively loading tasks. This makes the technique robust. Although other autonomic measurements like heart rate and skin resistance have also been found to provide similar information regarding sympathetic activity (and thus cognitive loading), pupillometry has been found to yield the most consistent and readily analysable results (Kahneman, Tursky, Shapiro, & Crider, 1969).

Pupillometry as a measure of cognitive load

In a seminal article, Hess and Polt (1964) found there to be a strong correlation between difficulty of arithmetic problems posed to participants and the magnitude of the increase in their pupil sizes. Further, they observed that after a question was asked, participants' pupils showed a gradual increase in diameter, reached a maximum size just prior to reporting an answer, and then reverted back to their original diameter shortly thereafter. In another article, Beatty and Kahneman (1966) built upon these original experiments and were able to

confirm two phases in the pupillary response to cognitive processing. First, they noted a loading phase with dilation corresponding to information gathering and an unloading phase where the pupil constricted as answers were verbalized by the participants. Based on the results from these studies as well as others, it became generally accepted that changes in pupil size reflect changes in cognitive processing load during task performance and provide information about processing resources. Specifically, more difficult cognitive tasks were found to cause both an increase in the amplitude and the latency of pupillary dilation (Beatty, 1982).

These early experiments were carried out with relatively onerous experimental processes that involved developing large quantities of photographs taken by cameras in precisely controlled environments and then manually measuring pupil size with a ruler. This made large-scale experiments impractical. Modern technology has allowed researchers to electronically collect pupil size data with stationary as well as mobile devices, obviating the need for time-consuming manual measurement and allowing for less stringent experimental environments. Some of the previously described studies that used arithmetic problems have now been replicated with the new technology, showing similar results. Figure 1 is taken from one of these studies that used a mobile eye-tracker to capture participant pupil size at a rate of 30Hz during arithmetic problem solving. As was first shown in the original experiments, the new technology also demonstrated that pupil size increased with increasing problem difficulty and changed predictably with phases of information gathering and delivery of responses (Szulewski, Fernando, Baylis, & Howes, 2014).

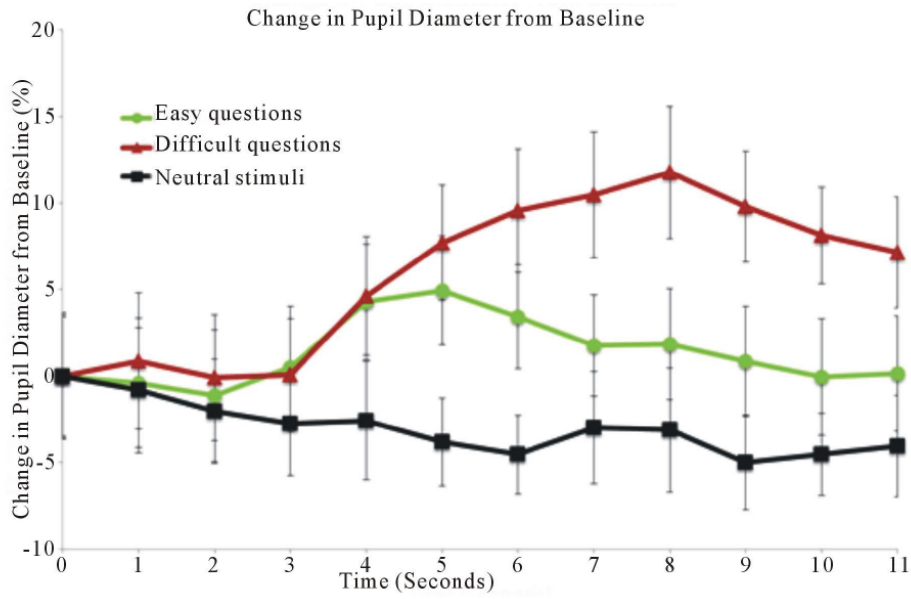


Figure 1*:* "Difficult questions resulted in peak dilation of 11.8% compared to baseline whereas "easy" questions resulted in peak dilation of 5.0% compared to baseline (p = 0.005). Time 0 to 3 seconds serves as baseline (3 seconds prior to question presentation); time 3 to 8 seconds corresponds to the time that the question was on the screen; time 8 to 11 seconds corresponds to the 3 seconds after the question was removed and the black dot appeared. [From Szulewski, A., Fernando, S. M., Baylis, J., & Howes, D. (2014). Increasing pupil size is associated with increasing cognitive processing demands: A pilot study using a mobile eye-tracking device. Open Journal of Emergency Medicine, 2014. Reprinted with permission].

The ease of use and precision of the newer technology has expanded the role of pupillometry to more theoretical realms. In addition to reliably demonstrating increased cognitive load with increasing question difficulty, pupillometry data have also shown that the modality of information presentation has cognitive loading effects. Using a remote eye tracker, Klingner, Tversky, and Hanrahan (2011) showed that cognitive load is higher for the same tasks when they are presented orally as opposed to visually. These experiments underscore the precision and expanded applications of this technique.

Other groups of researchers have also investigated the ability to measure cognitive load in novel environments using pupillometry. One such experiment by Palinko, Kun, Shyrokov, and Heeman (2010) investigated measuring mean pupil diameter change in drivers as they

operated a simulated vehicle while they were involved in simultaneous spoken dialogues. Pupil diameter changed as expected and the authors concluded that pupillometry was better in quantifying small changes in cognitive load in the simulator compared with other measures like lane position and steering wheel angle. Results from studies like this one suggest that pupillometry can be reliably used in more true-to-life situations in addition to well controlled laboratory settings. Importantly, during the driving simulator experiment, luminance varied only ± 5% in the simulated experimental environment which likely minimized the contribution of the light reflex to pupillary changes and allowed for a relatively clean signal. Changes in luminance become more of an experimental issue in real-world environments where background luminance varies to greater degrees.

On the whole, these studies seem to suggest that the construct being measured with pupillometry is cognitive load, although there is research to suggest that other factors (e.g. emotion, fatigue, age, pain and certain drugs) also contribute to change in pupil size (Holmqvist et al., 2011). Validity evidence for the use of pupillometry to measure cognitive load specifically was recently described by Szulewski, Gegenfurtner, Howes, Sivilotti, and van Merriënboer (2016). In this study, pupillometric measurements of cognitive load were compared to psychometric measurements of cognitive load across different question types, question difficulty and experience levels in a testing environment. Based on the predictability of the results and the strong correlation of the measurement instruments, the authors concluded that there is validity evidence to use either psychometric or pupillometric measurements to measure cognitive load in traditional testing environments.

Pupillometry research in medicine

Given the promising results of pupillary analysis in experimental settings and the increasing availability of the new technology, researchers have started to expand its use into other domains, including medicine. Medicine is a particularly interesting field in which to study cognitive load given its inherent characteristics where physicians regularly make high stakes decisions, often under considerable external pressures (including time and stress).

These characteristics are emphasized during non-routine emergency situations. One study that investigated critical incidents in the operating room examined anaesthesiology trainees' pupil sizes, among other physiological responses as surrogate markers of cognitive workload (Schulz et al., 2011). Participants' pupil sizes were found to increase as the severity of a critical incident increased. Although this pattern held true within scenarios, the authors found that there was no difference between sessions or individuals. This was thought to be due to individual pupil variations as well as external factors like lighting.

These issues raise the concern that external factors can skew pupillometric results and make it difficult to interpret the data reliably in real-world environments where luminance is not adequately controlled. All real-life physician-patient clinical encounters would, as a result, be affected. The main issue in these scenarios involves the light reflex which is capable of causing pupil diameter changes of up to 120% from baseline, which is far greater than the changes of up to 20% that can be attributed to cognitive processing demands. (Holmqvist et al., 2011; Laeng et al., 2012).

In an effort to mitigate the confounding effects of the light reflex, investigators often try to control for luminance during their experiments. Zheng, Jiang, and Atkins (2015) did just this and were able to confirm that pupil responses behaved as expected with changing sub-task difficulty in a simulated laparoscopic surgical experiment. In a related study, the same group noted that the rate of change of pupil size was better than pupil diameter in assessing mental workload of the simulated laparoscopic task (Jiang, Zheng, Tien, & Atkins, 2013).

To address some of these issues, new techniques (like the index of cognitive activity) have been designed to separate out the light reflex from pupil changes secondary to cognitive workload by measuring abrupt discontinuities in the pupil size signal (Marshall, 2002). This index of cognitive activity has been utilized in the objective assessment of surgical skill where pupil size (along with other eye and pupillary metrics) was used to objectively classify non-expert from expert surgeons in environments that were uncontrolled for luminance including a simulator as well as a live operating room (Richstone et al., 2010).

Pupillometry and expertise

Performance on tests has universally been utilized to measure the construct of ability, intelligence, competence or expertise in a domain. Despite its wide use, test-taking is known to have many limitations as a surrogate marker to measure these constructs. Applications of pupillometry have allowed researchers to delve deeper into this area than simply examining performance. This is particularly interesting when considering cognitive processing as subjects answer questions correctly. Based on the traditional view of assessment in test-taking, two individuals who get the same score on a test are thought to have equal domain-specific skill (or ability or expertise). The reality is more nuanced. Figure 2 is taken from a study by Ahern and Beatty (1979) which shows the cognitive processing demands (as measured by pupillometry) of participants with both high and low intelligence as defined by Scholastic Aptitude Test scores as they were faced with arithmetic problems (and answered these problems correctly). Based on traditional assessment modalities, both groups of individuals would be assessed equally for having answered correctly. A closer analysis, however, revealed that the group with "lower intelligence" had greater increases in pupillary dilation than the "higher intelligence" group at all question difficulty levels. Essentially, the group with the lower intelligence had to "think harder" to achieve the same correct response as the group with higher intelligence.

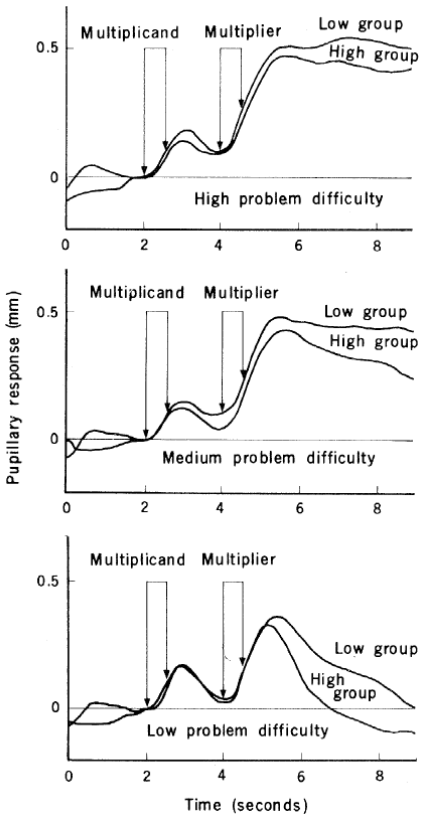


Figure 2*:* Averaged task-evoked pupillary responses for correctly solved problems at three levels of difficulty for subjects in the high and low groups of psychometrically measured intelligence. At all difficulty levels, larger pupillary responses are observed for subjects in the low group. [From Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*(4412), 1289-1292. Reprinted with permission from AAAS.]

Moving away from intelligence defined by standardized testing, a study by Szulewski et al. (2015) found similar results in novices and trained physicians as they answered clinically-based multiple choice questions. The participant groups in this study were divided not by intelligence, but by clinical experience. Those with more clinical experience (the trained physician group) had smaller changes in pupil diameter as they answered the questions compared to the more novice group when both groups answered correctly (See Figure 3). In another study, Tien et al. (2015) found that junior surgeons had greater pupil sizes than

expert surgeons during open inguinal hernia repair. Both of these studies (which divided physician participant groups based on experience level) emphasize that those with less experience expend more cognitive load than those with more experience when they perform domain-specific tasks, even when the measured outcome is the same.

Although it is reasonable to assume that these observed differences are due to different experience levels, one might argue that there may be other confounding factors between the groups that could skew the results. This is a potential issue for any cross-sectional study. A study by Richstone et al. (2010) suggests that it is in fact experience/expertise that is responsible for the pupillometric changes between groups, as opposed to another confounder. As part of their study, they examined one non-expert surgeon three times over the course of 18 months both in simulated and live surgical environments. During this longitudinal analysis, they found that it became increasingly difficult to differentiate this non-expert from the expert surgeon group as his pupil metrics became more expert-like over time with increased training and experience. This finding suggests that the differences between groups of participants are in fact due to skill or expertise, as opposed to another confounding factor. Overall, these studies suggest that there is empirical evidence that those with more domain-specific experience exhibit a certain cognitive efficiency as they perform tasks associated with their training and experience that their novice counterparts have not yet developed.

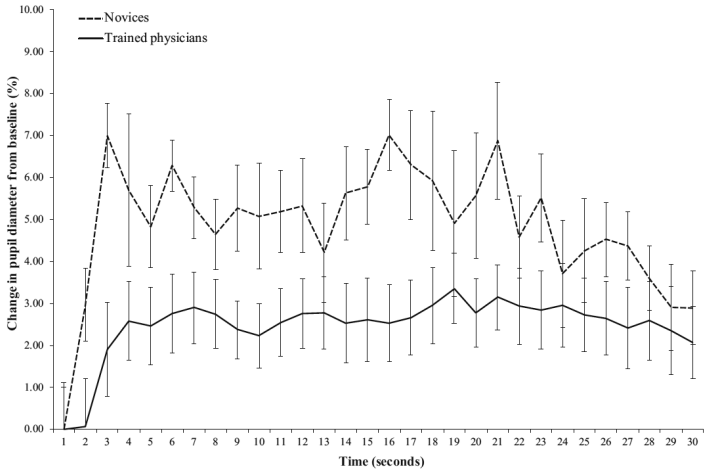


Figure 3*:* Results of an analysis of correctly answered clinical multiple-choice questions. The increase in the pupil diameter of novices was significantly greater than that of trained physicians ($P < 0.001$). [From Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987. Reprinted with permission from AAMC.]

It is debatable whether differences in cognitive efficiency are relevant during a test where a student is asked a sequence of questions and he/she generally focuses all of his/her working memory onto the question at hand before moving onto the next one. Moreover, it is debatable whether an assessor would even want to know this information. Arguably, however, this cognitive efficiency in the "more intelligent" or "more skilled" or "more experienced" or "more expert-like" group becomes relevant in complex situations with competing priorities, as the less cognitively strained individual will have a greater proportion of his/her working memory available for other cognitively demanding executive functions.

One such area where competing priorities often coexist and where cognitive efficiency might be beneficial is clinical medicine. During medical emergencies in particular, a physician team leader is cognitively tasked not only with making appropriate medical decisions but also employing crisis resource management techniques (like leadership skills, situational awareness, communication skills and resource utilization) to optimize patient care (Hicks, Bandiera, & Denny, 2008). It logically follows that cognitive efficiency in medical

decision-making will more readily allow the physician leader to perform these simultaneous crisis resource management tasks to a higher level given the real constraints of human working memory. Anecdotally, cognitive efficiency seems to evolve with experience. The "anatomy" of working memory is thought to change with the development of expertise and it is likely that certain clinical tasks cognitively load experts and novices in different ways (Szulewski et al., 2015). This evolution of the thinking process is tied to expertise development.

Typical experimental conditions for pupillometry studies

As outlined in this review, researchers have successfully used pupillometry as a cognitive load measurement tool in numerous experimental conditions. These range from relatively simple experiments where pupillometric data are gathered as participants are presented with written or verbal questions and are tasked to solve problems in fields including arithmetic and language, among others. Other experiments involve the use of different stimuli including photographs or even simulated driving environments. In medical applications, pupillometry has similarly been used in various settings including test-taking as well as more high fidelity environments like simulation and actual physician-patient clinical encounters. The task instructions provided to participants are equally variable and range from solving provided problems to performing operations in live surgical environments.

Constraints of pupillometry

Though it is clear that pupillometry provides useful information about both visual as well as non-visual stimuli, the technology has a number of constraints. Until relatively recently, accurate pupillometry studies required cumbersome experimental environments and tedious data collection and analysis. Although some of these issues have been addressed with new technology, the cost of this technology poses new financial barriers for certain researchers on smaller budgets. This is especially relevant for those part-time researchers who may want to incorporate pupillometry into their professional and teaching duties, like

academic physicians. This reality suggests that, for the time being given the costs, pupillometry research is more likely to occur at a theory-building level. As a result, some of its potential benefits in adjusting cognitive load on an individual learner and individualizing and optimizing education will remain elusive until the technology becomes cheaper and more readily available for teachers.

Another significant constraint of the technology relates to the accommodation and light reflexes. Although pupillometry provides consistent and fairly easy-to-interpret data in experimental conditions of constant ambient light and focus distance, data output in real-world conditions is suboptimal. As previously discussed in this review, the index of cognitive activity has been designed in an effort to overcome some of these obstacles. Although this technique allows for extraction of valuable information from large data sets with changing ambient light, the results are more coarse and provide less precise and detailed information about shorter-term cognitive changes that might be relevant in studying precise comparisons between groups performing shorter tasks (Klingner, Kumar, & Hanrahan, 2008). In addition, because this metric is a commercial product and its algorithm is not made publicly available, it cannot be replicated nor adequately studied. As a result, it is of limited benefit to researchers.

Another consideration in pupillometry research is participant age. Older individuals generally have pupils that are smaller and are more restricted in their ability to dilate compared to younger people (Holmqvist et al., 2011; Piquado, Isaacowitz, & Wingfield, 2010). Since many studies compare cognitive load between novices (who are usually younger) to experts (who are generally older), this might lead to confounding, as a smaller pupil diameter change may be due to a combination of increased age as well as decreased cognitive load. Cognitive load researchers should be aware of this issue and either control for participant age (where possible) or correct for it. Correction measures include expressing pupil size changes relative to a baseline measurement and/or age-adjusting for pupil size and reactivity based on participants' pupil responsiveness to a range of experimental light stimuli (Piquado et al., 2010).

Finally, the accuracy of pupillometric measurement is dependent to some degree on gaze position, with greater systematic error occurring when the eye is looking away from the eye-tracker's camera (Brisson et al., 2013). Different eye-tracking devices attempt to correct for

this error, but the accuracy of pupillometric measures suffers from variable quality under these conditions. This is especially relevant for researchers studying cognitive load where the participant's gaze may move away from centre.

Conclusion

Pupillometry is a robust and reliable method for studying cognitive load. Since its inception as a scientific field in the 1960's, it has evolved greatly. The development of new technology to measure pupil size that can electronically gather pupil data at high rates has led to the increased use of pupillometry in diverse fields. Despite the inherent constraints of the technique including interference by luminance and its cost, pupillometry remains a promising metric for researchers to utilize in the study of cognitive load as it can provide insights into the human thinking process that are otherwise unobservable. It has a particularly promising role in the field of medicine and in the study of physician expertise development. Utilizing pupillometry to better understand and optimize physician cognitive load (and overload) is clinically relevant and has the potential to directly impact medical education and ultimately patient care.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*(4412), 1289-1292.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*, 403-450.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science, 5*(10), 371-372.

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods, 45*(4), 1322-1331. doi:10.3758/s13428-013-0327-0

Burgess, P. W., & Simons, J. S. (2005). Theories of frontal lobe executive function: clinical application. In P. W. Halligan & D. T. Wade (Eds.), *Effecitveness of Rehabiliation for Cognitive Deficits* (pp. 211-231). New York: Oxford Univeristy Press.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*, 135-168. doi:10.1146/annurev-psych-113011-143750

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349-350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Hicks, C. M., Bandiera, G. W., & Denny, C. J. (2008). Building a Simulation-based Crisis Resource Management Course for Emergency Medicine, Phase 1: Results from an Interdisciplinary Needs Assessment Survey. *Academic Emergency Medicine, 15*(11), 1136-1143.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*: OUP Oxford.

Jiang, X., Zheng, B., Tien, G., & Atkins, M. (2013). Pupil response to precision in surgical task execution. *Studies in Health Technology and Informatics, 184*, 210.

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology, 79*(1p1), 164.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). *Measuring the task-evoked pupillary response with a remote eye tracker.* Paper presented at the Proceedings of the 2008 symposium on Eye tracking research & applications.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*(3), 323-332.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18-27.

Lang, G. K. (2015). *Ophthalmology*: Thieme.

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P., & van Merriënboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32-42.

Marshall, S. P. (2002). *The index of cognitive activity: Measuring cognitive workload.* Paper presented at the Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). *Estimating cognitive load using remote eye tracking in a driving simulator.* Paper presented at the Proceedings of the 2010 symposium on eye-tracking research & applications.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560-569. doi:10.1111/j.1469-8986.2009.00947.x

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

Schulz, C., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., . . . Schneider, G. (2011). Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. *British Journal of Anaesthesia, 106*(1), 44-50.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138.

Szulewski, A., Fernando, S. M., Baylis, J., & Howes, D. (2014). Increasing pupil size is associated with increasing cognitive processing demands: A pilot study using a mobile eye-tracking device. *Open Journal of Emergency Medicine, 2014*.

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L. A., & van Merriënboer, J. J. G. (2016). Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 1-18. doi:10.1007/s10459-016-9725-2

Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987.

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G.-Z., & Darzi, A. (2015). Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surgical Endoscopy, 29*(2), 405-413.

Whipple, B., Ogden, G., & Komisaruk, B. R. (1992). Physiological correlates of imagery-induced orgasm in women. *Archives of Sexual Behavior, 21*(2), 121-133.

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: Implications for medical education: AMEE guide no. 86. *Medical Teacher, 36*(5), 371-384.

Zheng, B., Jiang, X., & Atkins, M. S. (2015). Detection of Changes in Surgical Difficulty: Evidence From Pupil Responses. *Surgical Innovation, 22*(6), 629-635. doi:10.1177/1553350615573582

Appendix B

Decision-making During Medical Crises

Published as: Szulewski, A., Brindley, P. G., & van Merriënboer, J. J. G. (2017). Decision making in acute care medicine. In *Optimizing Crisis Resource Management to Improve Patient Safety and Team Performance* (pp. 13 - 20): Royal College of Physicians and Surgeons of Canada.

Introduction

Decision-making is fundamental to the provision of effective medical care. Early in training, health care practitioners (HCPs) are taught a linear-approach to decision-making that works well for the majority of stable patients. This begins by obtaining a patient's history, followed by performing a physical examination, then developing a differential diagnosis, next ordering investigations, and, finally, instituting therapy. For stable patients, this approach maximizes information-gathering and provides time for contemplation. In contrast, during medical crises, this strategy is impractical and potentially dangerous: especially if it delays time-dependant resuscitation. The provision of emergency care may be challenging for HCPs, and potentially perilous for patients.

During a medical crisis, the goal is to maximize patient-stability and minimize delays. Diagnosis and therapy should occur concurrently, and, importantly, at the expense of diagnostic precision. Data-gathering focuses more on what is immediately-available (i.e. vital signs and point-of-care analysis) and less on waiting for diagnostic tests (CAT scans; laboratory results). Similarly, consults are limited to specific interventions (e.g. intubation; surgery; help with resuscitation) rather than diagnostic opinions. To manage the patient-in-peril, the team needs to rapidly convert available data (i.e. an increasing heart rate) into usable information (i.e. the patient is worsening) and followed by a logical response (i.e. bolus fluids). The art of acute care medicine is ensuring that while we do not intervene without sufficient thought, we do not allow uncertainty to cause potentially harmful delays.

As outlined, the concurrent approach employed during crises downplays the need to establish an immediate etiologic diagnosis (e.g. streptococcal septicemia). Instead we often redefine uncertainty by providing broader temporary physiologic or pathophysiologic diagnoses (e.g. hypotension or septic shock). Missing diagnostic details and treatment gaps are filled in later when the medical crisis has abated and when traditional sequential decision-making strategies can be safely employed again. The concurrent approach increases the chance that the physician-leader and medical team can stay ahead of a rapidly evolving situation, and can simultaneously manage competing priorities.

Beyond the challenges of time-sensitive decision-making, the good doctor-leader must also maximize the effectiveness of the whole team, and despite high stimulus-density and high

clinical stakes (Hicks, Bandiera, & Denny, 2008). This can be done by employing well-established crisis resource management (CRM) principles. These CRM skills are reviewed in other sections of this book and include leadership and followership, situational awareness, communication skills, resource utilization and teamwork. Specifically, in this chapter we focus on the theory and practise of effective decision-making as well as the effect that experience, cognitive-load and working-memory have on decision-making itself.

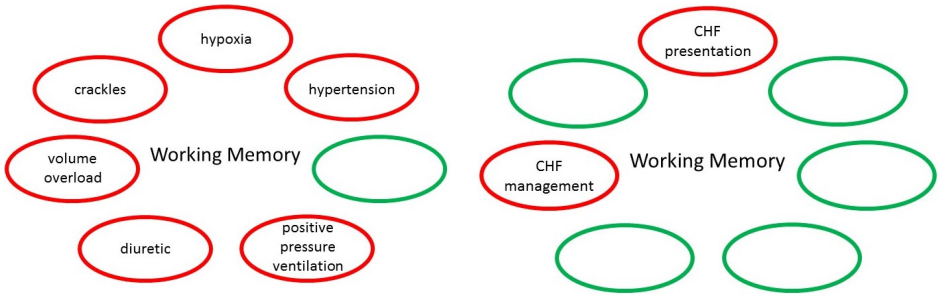
The Fundamentals of Medical Decision-Making

As outlined, decision-making, of one form or another, is important for all HCPs. However, it is central to medical doctors, who make the majority of high stakes decisions. Despite its importance, it is rarely deliberately addressed in traditional medical curricula. Instead, doctors typically gain most of their experience "on-the-job" during clinical work. With experience, most eventually become capable decision-makers; however, the process of decision-making, and deliberate strategies to optimize it, are often not fully appreciated by the decision-makers themselves.[2-4] In other words, clinicians often become unconsciously competent decision-makers.

Over a career, medical decision-makers hone their "intuition" and "clinical reflexes". However, it is often difficult for HCPs to articulate how or why they make particular decisions (G. Klein, Calderwood, & Macgregor, 1989). For example, an experienced physician can quickly identify the deteriorating asthma patient, decide to intubate, and institute appropriate therapy. When asked later what made him/her intervene so quickly, answers might include "the patient was fatiguing" or "if I hadn't then the patient was going to arrest". Though true, these judgements are intuitive (or intrinsically tacit) and difficult to relate to for novices. This often makes decision-making difficult to teach.

Understanding decision-making during crises involves addressing the limits of human working memory. We can only reliably manage a finite number of discrete elements of information (approximately seven), and an even smaller number when information-processing is required (Miller, 1956; Sweller, Van Merrienboer, & Paas, 1998). For example, for the novice who is managing a patient with congestive heart failure (CHF), these

information elements may be as basic as “hypoxemia”, “hypertension”, “crackles”, “volume overload”, “diuretic” and “positive pressure ventilation”. These six items approach the novice’s working memory capacity. In contrast, for the expert, multiple elements can be integrated into information “units” or “chunks”, such as “CHF presentation” and “CHF management”. This leaves a larger proportion of working memory available for other tasks. Figure 1 summarizes this concept.



A. **B.**

Figure 1: Working memory of a novice (panel A) and expert (panel B) treating the same patient with congestive heart failure. Ovals represent discrete elements of working memory possessed by an average individual (seven). Red ovals represent occupied components; green ovals represent available working memory, which can be activated for other tasks. The inclusive “chunked” information in the expert’s working memory allows him/her to simultaneously address other tasks, despite a physiologically similar working memory to the novice.

Models of decision making

Decision-making (also called problem-solving in some CRM models) is a complex topic. However, it has been summarized using various theoretical models from several professional domains. Two of these models – Gary Klein’s Recognition-Primed Decision (RPD) Making Model (G. Klein, 2008) and Daniel Kahneman’s Dual Process Model (DPM)

(Kahneman, 2003) provide a foundational understanding of the cognitive processes employed by expert HCPs.

Recognition Primed Decision-Making Model

The RPD model helps explain how successful decision-making can occur in complex, ever-changing, medical environments – and despite the constraints of human working-memory. As outlined, most experienced doctors, when faced with a crisis, do not consciously compare a multitude of options prior to acting. For example, an experienced physician managing an intubated trauma patient with hypotension and hypoxemia might expedite a lung ultrasound, see evidence suggesting pneumothorax, and rapidly decompresses the chest. This occurs rapidly not because that physician possesses special knowledge but rather because they are attuned to the possibility of tension pneumothorax in all patients with chest trauma, and cognizant of the danger of missing this diagnosis. He or she also pattern-recognizes the association between tension pneumothorax, positive-pressure ventilation, hypoxemia and hypotension.

In this way we can begin to define what makes an effective acute care doctor. He/she is one who can quickly focus on high-yield diagnostic clues, rapidly confirm their suspicions, address the key dangers, act expeditiously, and avoid wasting cognitive resources on extraneous details (Kahneman & Klein, 2009). Moreover, the experienced doctor is able to recognize when his/her initial course of action is flawed and is cognitively dextrous and confident enough to modify his/her response. If the therapeutic plan cannot be easily modified, then the next most plausible course of action is rapidly selected. This process is then repeated until an acceptable way forward is found (G. A. Klein, 1999). This sequence of steps forms the basis for the RPD model. Once again, it is in contrast to the traditional approach of linear information gathering and exhaustive hypothesis generation.

The recognition displayed by the expert physician is analogous to intuition and is central to RPD. A junior doctor may not immediately recognize the previously described cluster of signs and symptoms as typical of a tension pneumothorax. As a result, decision-making becomes more analytical and time-consuming for the novice. Despite every good intention,

patients can suffer the consequences of delayed decision-making in a time-sensitive situation.

Dual-process model

An alternative to the RPD model is the DPM model of decision-making described by Daniel Kahneman. This model conceptualizes thinking and decision-making into System I and System II. System I is involved in intuitive judgments that are fast and automatic. These judgments are relatively effortless and lack a sense of voluntary control (Kahneman, 2011). For example, the experienced clinician who enters a ward and declares within seconds that a patient is "sick" or "not sick" is using System I. Of note, the ability to simply recognize a 'sick patient' is every bit as important as learning factual knowledge or mastering manual skills. The key is that these are familiar situations and therefore the physician recognizes a pattern. As such, adept decision-making requires repeated and regular exposure.

System II is slower and more logical. It is activated when a situation is unfamiliar and therefore deviates from the world-model that System I maintains. It replaces fast and relatively *effortless* intuition with *effortful* logical reasoning (Kahneman, 2011). For example, the patient with resistant hypotension eventually found to have adrenal insufficiency is likely to have induced a physician's System II processing. The ability to step-back from a crisis and use System II reasoning during the stress of resuscitation is another hallmark of the experienced and effective HCP. Again, this requires regular and repeated exposure.

Complementary models

The DPM is supported by the psychological literature regarding cognitive errors and biases, whereas the RPD model is supported by expert-intuition and decision-making theory. However, these approaches overlap, and are better thought of as complementary rather than oppositional. For example, the intuition that informs the RPD model is similar to System I processing within the DPM. Recognition (or intuition, or System I processing) is relatively accurate in experts' hands but potentially problematic for novices. The danger of inexperience in the novice, or fatigue in the expert, is that both could oversimplify (or

morph) complex medical problems in order to fit a previous (different) encounter. This cognitive bias is referred to as the simplifying heuristic (Kahneman & Klein, 2009).

In Situ Decision Making

Recognition, expertise and cognitive load

Recognition is key to clinical decision-making. What is less clear is how HCPs develop expertise in recognition. Research suggests that the practice environment needs to provide both sufficient valid cues as well as the opportunity to identify these cues (Kahneman & Klein, 2009). Accordingly, chaos, distraction and unhelpful team-mates can affect the likelihood of timely recognition. Also, as outlined, for HCPs to become skilled in resuscitation medicine and make effective decisions in crisis situations, they need to have sufficient exposure and experience. This can be gained through clinical encounters or well-crafted medical simulations. It is unfair and illogical to expect HCPs who are not exposed to regular decision-making during crisis to perform at a high level during such a rarely encountered situation.

Experiments based around Operating Room emergencies, and managed by Anesthesia residents, have suggested that resident physicians exhibit one of four problem solving approaches (Rudolph, Morrison, & Carroll, 2009). Residents who are *"*stalled*"* find it difficult to generate diagnostic possibilities or coordinate their response. Others are "fixated" and quickly generate a plausible but incorrect diagnosis, and have trouble deviating despite alternate cues (so called "premature closure"). "Diagnostic vagabonds" produce a large number of possibilities but fail to rule them in or rule them out. The "adaptive" group is the most effective. These residents generate a number of plausible diagnoses, rule certain ones out, and respond appropriately.

As HCPs gain experience and develop expertise, they are more likely to recognize immediate threats and rapidly intervene, even at the expense of diagnostic accuracy. In fact, cognitive processing and decision-making strategies should naturally mature over time, and this process is one way to define clinical expertise. (Szulewski, Roth, & Howes, 2015) Experience decreases cognitive load and frees up more working memory and higher-level thinking. In

contrast, the inexperienced physician may be too cognitively overloaded to consider alternate diagnoses or pre-emptive interventions. In this way, cognitive overload can erode, and de-prioritize, the novice's CRM skills. (Figure 2 provides a graphical representation of some of these concepts).

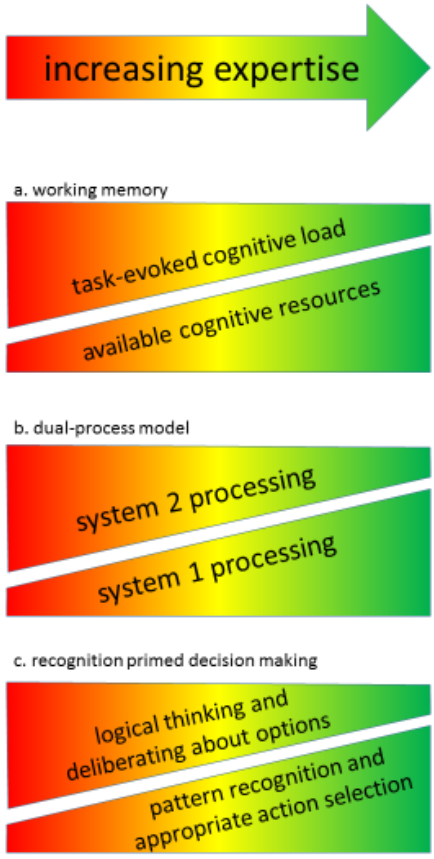


Figure 2: Relationship between increasing expertise, working memory, dual-process model and recognition primed decision-making.

Teaching decision making

If we accept that experienced physicians, bolstered by regular clinical exposure, are effective crisis decision-makers, then it makes sense to teach the RPD model. Accordingly, Cohen and Freeman (1997) have used this model to address critical-thinking using real-life clinical cases. In order for teaching to be effective, clinical information should be presented in an unpredictable sequence (random practice schedule). This not only mirrors acute care but also forces learners to critically compare and contrast new data with whatever came before (A. S. Helsdingen, T. Van Gog, & J. J. van Merriënboer, 2011). For novices it might be necessary to simplify the cases and provide guidance (or cognitive nudges) that help them recognize what is most relevant, and what is most distracting (Van Merriënboer, 1997). During instruction, learners should focus on four beneficial activities. These are (1) *creating a story*: where all existing evidence is incorporated and explained, and where reasonable assumptions are made when there is uncertainty, (2) *testing a story*: where inconsistencies and uncertainties are identified and the story refined through deliberate testing, (3) *evaluating a story*: where plausibility is questioned by playing the devil's advocate, and (4) *quick testing*: where the time available and the consequences of actions are pre-determined, thereby encouraging more immediate action if delays are unacceptable (A. Helsdingen, T. Van Gog, & J. Van Merriënboer, 2011).

Effective instruction in critical decision-making requires a pre-brief that describes a cognitive strategy that can steer the decision-making process. It also requires a skilled facilitator who can prompt the learner to self-reflect on his/her developing strategy so that it can be refined. Prompts should help learners prevent mistakes, challenge their biases, and ensure they remain open to other explanations. When learning situations are presented in an unpredictable sequence, the use of retrospective prompts (e.g., were there any similarities between the last two situations?) are more effective then proactive prompts (e.g., are there any similarities between the following two situations?). The combination of random practice schedule and retrospective prompts increases the likelihood that skills are transferred from one situation to the next.[15] In this way, education around decision-making can benefit both practitioners and patients.

Summary/Conclusion

Effective decision-making is a complex but essential skill for those charged with caring for acutely ill patients. Despite a substantial body of knowledge about decision-making in non-medical domains, it is rarely taught or coached in medical training programs. Fortunately, there is emerging evidence surrounding decision-making that can be readily adapted to acute care medicine. The goal of this chapter is to allow HCPs to better understand their own decision-making habits and how this impacts crisis resource management in everyday clinical practise.

Acknowledgments

Drs. Marco Sivilotti and Bob McGraw for their thoughtful input.

References

Cohen, M. S., & Freeman, J. T. (1997). Understanding and enhancing critical thinking in recognition-based decision making. In R. Flin & L. Martin (Eds.), *Decision making under stress: Emerging themes and applications.* Brookfield, VT: Avebury Aviation.

Helsdingen, A., Van Gog, T., & Van Merriënboer, J. (2011). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. *Journal of Educational Psychology, 103*(2), 383.

Helsdingen, A. S., Van Gog, T., & van Merriënboer, J. J. (2011). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction, 21*(1), 126-136.

Hicks, C. M., Bandiera, G. W., & Denny, C. J. (2008). Building a Simulation-based Crisis Resource Management Course for Emergency Medicine, Phase 1: Results from an Interdisciplinary Needs Assessment Survey. *Academic Emergency Medicine, 15*(11), 1136-1143.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 1449-1475.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515.

Klein, G. (2008). Naturalistic decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 456-460.

Klein, G., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. *Systems, Man and Cybernetics, IEEE Transactions on, 19*(3), 462-472.

Klein, G. A. (1999). *Sources of power: How people make decisions*: MIT press.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.

Rudolph, J. W., Morrison, J. B., & Carroll, J. S. (2009). The dynamics of action-oriented problem solving: Linking interpretation and choice. *Academy of Management Review, 34*(4), 733-756.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987.

Van Merriënboer, J. J. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training*: Educational Technology.

Appendix C

Through the learner's lens: eye-tracking augmented debriefing in medical simulation

Published as: Szulewski, A., Braund, H., Egan, R., Hall, A. K., Dagnone, J. D., Gegenfurtner, A., & van Merriënboer, J. J. G. (2018). Through the Learner's Lens: Eye-Tracking Augmented Debriefing in Medical Simulation. *Journal of Graduate Medical Education*, 10(3), 340-341.

Setting and Problem

Effective debriefing is a critical component of simulation-based education, and experts agree that a significant proportion of learning in simulation occurs during the debrief. Traditional debriefing techniques like the Plus Delta or Advocacy/Inquiry approaches are useful, but they may be limited if the full picture of the decision-making processes that occurred during the simulation is not known. Memory science suggests that it is unlikely that residents can accurately and consistently recall specific situational cues that precede clinical decision-making. For this reason, the rationale for clinical decision-making may be difficult to unpack and may be susceptible to recall biases (eg, hindsight bias). Although bird's eye video could provide context, it too does not allow for visualization of the clinical context from the resident's perspective. To get inside the head of a trainee and to provide more targeted feedback, we propose a novel method of cued retrospective debriefing augmented by eye-tracking technology.

Intervention

Emergency medicine residents were outfitted with mobile eye-tracking glasses during 2 resuscitation-based objective structured clinical examinations (OSCEs) in a high-fidelity patient simulation laboratory. Twenty-two residents participated over the course of 2 examinations; 13 residents completed both examinations. The eye-tracking glasses recorded the simulated encounters from a first-person perspective while measuring pupil position at 50Hz. This allowed for the creation of a video that showed the simulated encounters from the resident's perspective, with a superimposed gaze indicator that displayed where the resident was looking at all times.

After the completion of the OSCE, residents participated in a traditional faculty-led debrief. Immediately afterward, each resident met with a separate faculty debriefer who led an individualized debrief that was augmented by review of the residents' own eye-tracking videos. The debriefer played the videos back to each resident, pausing them at critical junctures and inquiring about each resident's decision-making and crisis resource management decisions. Individualized feedback, which invited the resident to compare and

contrast his/her decision-making with the decision-making of an expert, was provided based on the ensuing discussion.

Each resident was then asked to compare the utility of the traditional debriefing technique to the novel debriefing technique with an audio-recorded semistructured interview. Over the course of 2 OSCEs, interviews from 35 individual debriefings (each based on two 10-minute simulation scenarios) were analyzed using a thematic emergent design through a phenomenological lens.

This study received approval from the Queen's University Research Ethics Board; all participants provided informed consent.

Outcomes to Date

Three main themes emerged from the qualitative analysis. See Table 1 for a list of themes and example quotes.

All residents reported that the cued retrospective debriefing augmented by eye-tracking was useful for their learning. This novel approach encouraged residents to reflect on their performance, leading them to critique responses to specific situational cues and to identify new insights into their performance. Many residents reported that these insights were missed during the initial traditional debrief and only came to light by virtue of the eye-tracking video review.

Though primarily used in the research realm because of financial constraints, the cost of eye-tracking technology is quickly decreasing. This will allow medical educators to harness the benefits of this technology in simulation debriefing where it has the potential to enrich the process of learning.

Table 1: Summary of Emergent Themes

Theme	Example Quote
New insights and reflective thinking	"I think it's a good way to do a lot of self-reflection, a lot of ways to do real-time assessment. Like, I can see what I was thinking at the same time." (P19)
Identifying errors from video review	"Yeah, I sort of saw where I got distracted and, and led off-course better than I think I would've otherwise, had we not debriefed and seen where I started to focus on things that were perhaps, more distracting than anything else." (P11)
Added value of eye-tracking technology	"And I love good feedback and different molds of doing this, and this is an incredible tool. Like I've never seen the ability to see where somebody's looking . . . it's an amazing tool for identifying cognitive overload." (P18)