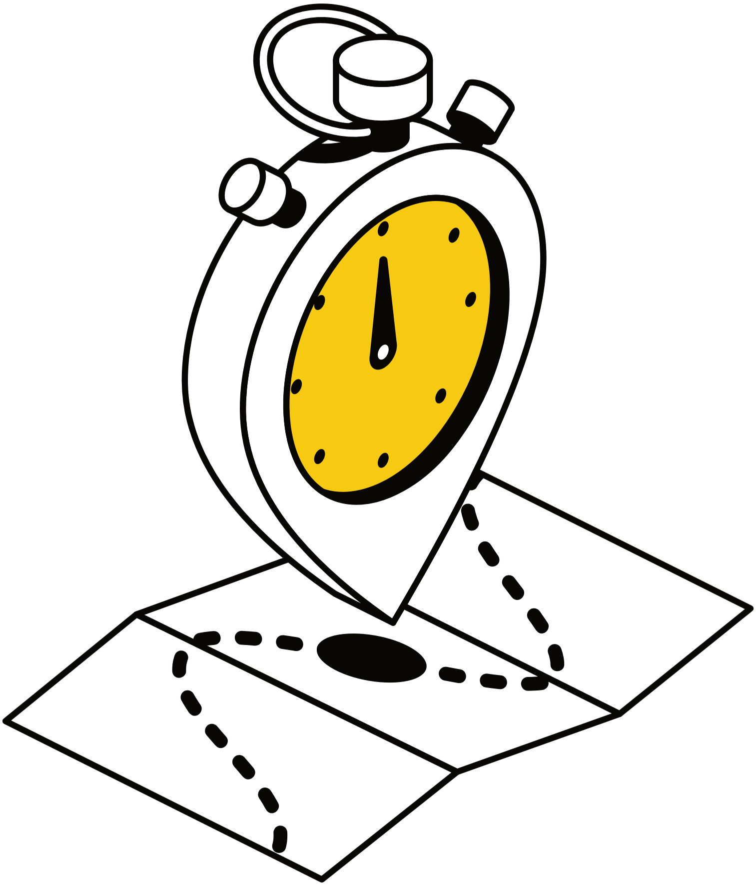
The medical pause in simulation training

The Medical Pause in Simulation Training



Joy Yeonjoo Lee

The Medical Pause in Simulation Training

Joy Yeonjoo Lee

The Medical Pause in Simulation Training

The research reported here was carried out at

Maastricht University | Maastricht UMC+

In the School of Health Professions Education

School of Health Professions Education

In the context of the research school

ico

(Interuniversity Center for Educational Research)
and was funded by the Netherlands Organization for Scientific Research (055.16.117)

NWO Netherlands Organisation for Scientific Research

Paperback ISBN: 978-94-6458-275-8
Ebook ISBN: 978-94-6458-272-7
NUR 840

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and design: Erwin Timmerman, persoonlijkproefschrift.nl

The Medical Pause in Simulation Training

DISSERTATION
to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus
Prof. dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday 8th of June 2022 at 10:00 hours

by

Joy Yeonjoo Lee

Supervisor(s):
Prof. dr. J.J.G. van Merriënboer
Prof. dr. H. Jarodzka, Open Universiteit, Heerlen

Co-supervisor:
Dr. H. H. L. M. Donkers

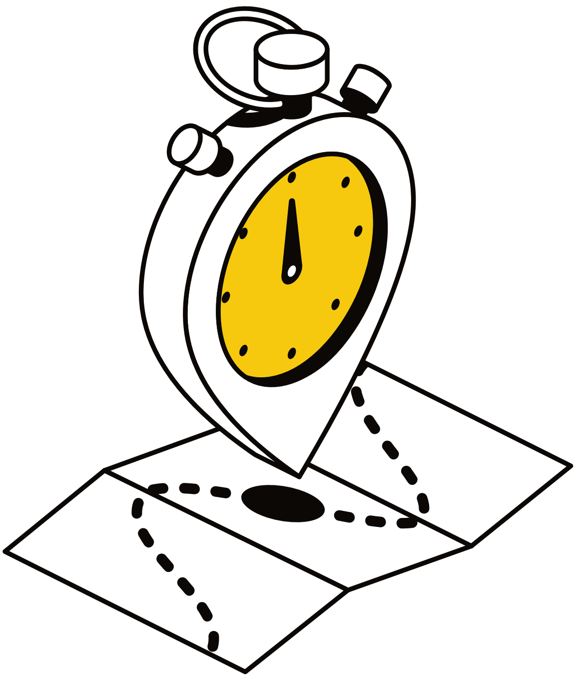
Assessment Committee:
Prof. dr. A.B.H. de Bruin (Chair)
Dr. M. Asoodar
Dr. M. Dankbaar, Erasmus University Medical Center, Rotterdam
Prof. dr. A. Gultz, Lund University, Sweden
Prof. dr. W. N. K. A. van Mook

Contents

"Mom, one does not simply pause an online game"

Chapter 1

General Introduction



"To err is human": As in many critical fields such as aviation, nuclear power, and chemical industry, human errors are one of the leading causes of advert events in healthcare.[1,2] To minimize human error and improve patient safety, diverse strategies for performance and quality improvement in healthcare systems have been attempted.[3] Among these strategies, the timeout, where the current action is paused to secure safe measures, has been adopted in various settings.[4,5] Research has demonstrated that the application of timeouts significantly decreases patient morbidity and mortality and enhances healthcare delivery performance.[6-8]

Although timeouts in general refer to formal team practice (e.g., surgical timeout, diagnostic timeout),[9,10] researchers have called for more focus on informal and individual timeouts (i.e., pausing) that can be applied to a broader range of practice, as recent medical interventions become more dynamic and fast-paced. Coined as the "reflective pause", "the mindful surgeon", or "slowing down when you should",[11-13] discussions on pausing have been cultivated in different contexts to spur cultural changes in healthcare systems.[14] *The medical pause* is a novel term developed through the course of the project reported in this PhD thesis, defined as a formal or informal decision to stop current action in medical procedures for a brief time to perform beneficial cognitive activities that enhance performance and patient safety.[15]

However, the pausing practice in any form appears to be not fully acknowledged both in workplaces and education. Reportedly, compliance with timeout protocols is so low that it dilutes the clinical benefits of timeouts.[5,16,17] Healthcare professionals are often unwilling to interrupt current procedures by taking pauses, as it seems to be inefficient.[5,17] Educators in training programs do not sufficiently consider pausing as a professional skill, rather focusing on algorithmic task approaches that hinder the development of critical thinking.[18,19] Students and trainees tend to disregard the benefits of pausing and rush to perform given tasks.[20,21] In research, studies on pausing have been developed in divergent contexts, which hampers the development of a coherent research framework and the conceptualization of pausing as a professional skill.[15,22] Moreover, these studies heavily rely on post-action self-repor for data collection. Apart from its common drawbacks (e.g., bias, forgetting, fabrication), self-reporting is not suitable to capture specific behaviors that arise during performance, calling for objective real-time measures to support research on pausing.[23]

The studies reported in this PhD thesis aim to foster the understanding of the medical pause as a professional skill that should be taught in educational programs, by investigating the effects of pausing both theoretically and empirically. The thesis contributes to the literature in three ways: First, I develop an integral framework to understand the medical pause by integrating basic concepts for safety culture developed in different contexts. Cognitive mechanisms shared between them are recognized based on concepts from educational psychology such as *cognitive load*[24,25] and *self-regulation*.[26] Second, moving from the theoretical grounds to educational translation, practical strategies to train pausing skills within the medical curriculum are suggested. Useful strategies such as *scaffolding* and *cognitive and metacognitive aids* are introduced, by employing educational theories of complex learning.[27,28] Third, the effects of pausing on different aspects of performance and learning are empirically examined, by using computer-based simulations (e.g., serious games, VR simulation) and objective measurement methods such as eye-tracking and game-log analysis.

In the following, I review previous research on pausing skills in medical education. Next, cognitive load theory is introduced to explain the two major cognitive processes that constitute pausing skills: relaxation and reflection. Computer-based simulation, then, is described as a promising environment for pausing skills for both educational and research purposes. Lastly, I present an overview of the five studies that build up this thesis.

1. Pausing Skills in Medical Education

The importance of pausing should be stressed especially in medical education for three reasons. First, tasks in medicine are often characterized as irreversible, associated with finality of action (e.g., harm or death of a patient).[23,29] This requires practitioners to be particularly circumspect and mindful by using safety measures such as timeouts. Second, medical tasks are highly uncertain and event-driven: Important information or unexpected events can emerge during task performance (e.g., finding abnormal anatomy during a surgery). This requires practitioners to continuously adapt their performance to changing environments by pausing the current procedure and redirecting it.[23,29] Third, beneficial cognitive activities during pauses (e.g., reflection) can facilitate continuous learning.[15,30] Since medical tasks are context-dependent and lengthy (e.g., several hours of surgery) and rare scenarios are difficult to replicate for practice purposes, creation of learning opportunities just-in-time in the context of actual performance is crucial.

Despite these benefits, traditionally, not many studies have paid attention to pausing (reflection *during* tasks), while there have been a large number of studies on debriefing or post-action reviews (reflection *after* tasks).[23] However, as the standards of patient safety have increased recently,[1,2] more studies on pausing have appeared in different contexts. In team performance, forced pauses initiated by formal protocols are studied, contributing to the development of surgical timeouts,[7,31] handoff protocols such as SBAR[32] and I-PASS,[33] and checklists.[34-36] Most of these studies reported positive effects of pausing on improved performance and reduced adverse events including morbidity and mortality.[37,38] Some researchers suggested to expand the standard surgical timeout into a comprehensive concept of "preparatory pause" or "preoperative briefing", and verified its positive effects on performance.[7,31]

Moulton et al. have broadened the concept of pausing to informal and individual practice, introducing the "slow down when you should" phenomenon.[22,39] They defined this phenomenon as the transition from an automatic to an effortful cognitive processing mode.[22] "Stopping" was identified as the most extreme example among other slowing down moments such as "focusing more intently" or "fine-tuning".[40] Recognizing the slowing down as the hallmark of expertise that allows for creative solutions and adaptive performance, they argued this practice should be taught in training programs to cultivate "mindful surgeons"[12,13] Some researchers also supported the idea of promoting pausing skills in training programs, for instance, integration of timeout practice in simulation-based team training.[18,19]

In comparison, in education in general, pausing has been studied mostly in the context of multimedia learning or instructional animations. As in medical training, students reportedly tend to be unaware of the importance of pausing during the tasks [41,42] and educational support for pausing is lacking.[43,44] Concerning the effects of pausing on learning, research has presented inconsistent results, [21,45-49] and no clear consensus exists about the definition of pausing and when and how to prompt it.[50-52] Researchers have pointed out that a systematic understanding of pausing is required to further the research,[53,54] and that, to spur this understanding, real-time objective measures should be leveraged to unveil the actual effects of pausing on learning.[53,55,56] Although some findings from general education are worth referring to, I consider they may not be directly applicable to the medical pause. Noticeably, the task environments used in these studies are linear and passive (e.g., learning from texts on the screen), while those for medical training are dynamic and the learners' interaction with the environment alters the consequences even more.

2. Relaxation and Reflection

Cognitive load theory (CLT) provides a powerful framework to explain learning processes in complex professional domains.[24,25] CLT posits that diverse mental processes impose different types of cognitive load on working memory. While discussions about these types are still continuing within the CLT community,[57,58] I suggest a new typology with three types of cognitive load to better explain pausing processes in medical practice: primary load (PL) imposed by domain-specific task performance (e.g., a surgical task), secondary load (SL) from domain-general processes that support the primary performance (e.g., self-monitoring or situation awareness [59,60]), and extraneous load (EL) that does not contribute to the performance (e.g., distracting conversation during the surgery). SL is particularly important in pausing skills, as it pertains to self-monitoring in uncertain and event-driven task environments to decide when and how to take pauses. Similar to the eustress status in the stress literature,[61,62] to maximize quality performance, working memory should be neither overloaded nor underloaded. In other words, the total level of cognitive load should be titrated below working memory capacity, while still maintaining positive cognitive processes (i.e., sources of PL and SL) in an active state. I argue that pausing is the professional skill to find and maintain this equilibrium via two types of cognitive activities during pauses: *relaxation* that discharges overload in working memory, and *reflection* that boosts cognitive activity by restructuring schemas in long-term memory (i.e., learning).

Figure 1. Transitions between different cognitive load states

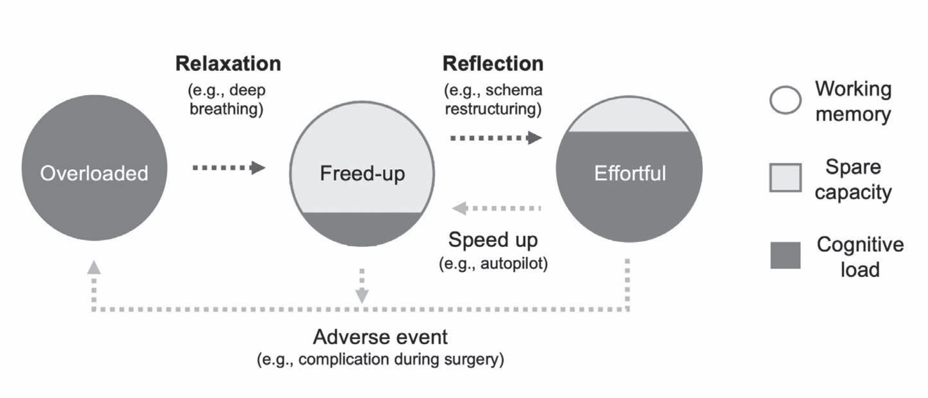


Figure 1 describes how various factors including relaxation and reflection trigger transitions between three different cognitive states: overloaded, freed-up, and effortful. The state of *overloaded* is caused by adverse events such as time pressure and complications, resulting in acute stress and heightened emotions.[25] Relaxation processes (e.g., deep breathing) transition this state to the *freed-up* state, through management of the source of overload or cognitive rest that recovers depleted working memory.[63] This freed-up state is a transient phase with spare capacity created in working memory by either relaxation or speeding-up (i.e., automatic performance of routine task aspects). Reflection activates positive cognitive processes to transition the freed-up to the *effortful* state. In this state, working memory is ideally optimized to keep control over performance and facilitate learning.

For example, in a difficult resuscitation case, a physician may pause the procedure to calm down (relaxation) and summarize the current situation by reviewing previous performance, asking for peers' opinions, and planning next actions (reflection). Note that these processes rely on accumulated experience and finely honed skills. Reportedly, experienced physicians are better skilled in those strategies to cope with intense situations.[64] They have developed clinical judgment and time management skills to assess the time available and the risk of using the time,[8,65] which requires years of experience and training.

3. Computer-Based Simulation

While real task environments can be suitable for some cases in medical training, simulated task environments can be more beneficial for educational purposes.[27] They can offer a wide spectrum of scenarios including rare and critical cases, expanded access to authentic task environments, and safer environments that do not risk patient safety or loss of valuable materials. More importantly, they allow for control over learning content (e.g., rearranging task sequence, providing guidance) and performance speed to pace task progress (e.g., pausing and resuming performance). This flexibility is important for education as instructors can tailor their level of support by adapting the task progression of learning tasks to learners' competency levels.

Computer-based simulation (CBS) is a technology-enhanced learning environment which makes the benefits of simulated task environments even stronger. Rare cases can be easily designed by programming patient physiology, and simulations can become more accessible through online distribution. Diverse CBS formats, such as web-based multimedia, serious games, and virtual reality, have successfully replicated clinical settings.[66-68] Moreover, it provides a promising way of training pausing skills by allowing for easy installment of pause functions. By pressing a pause button and deliberately taking a pause, learners can practice when and how to take pauses in different scenarios.

Another gain from using CBS is that it yields objective behavioral data during performance in real time (e.g., log data in serious games). These data sets can be even more informative when integrated with physiological data such as eye movements. Eye-tracking is a powerful information resource especially for CBS, since CBS heavily depends on visual stimuli. It provides insights on diverse cognitive processes during human-computer interaction.[69] For instance, the transition rate (i.e., eye gaze allocation across different objects) can reveal the level of situation awareness or vigilance.[70,71] Pupillometry can demonstrate the level of cognitive load (increasing pupil size is associated with higher cognitive load), allowing for expertise assessment.[64,72]

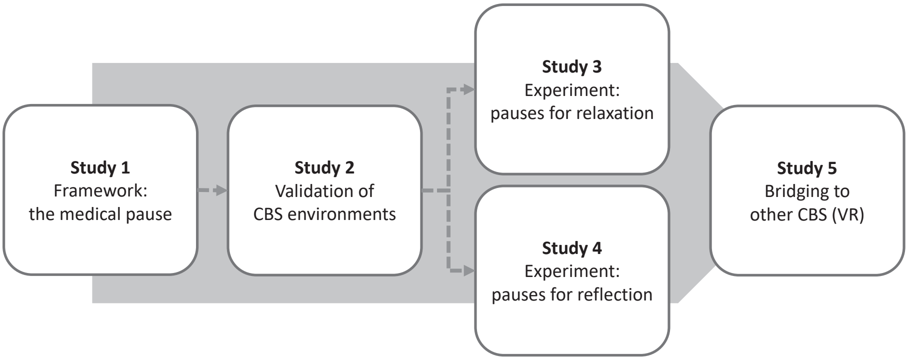
The availability of real-time data is invaluable for both research and education purposes. The data sets allow for a deeper understanding of cognitive processes in pausing. Educators and learners can use the data as cues for decision making on when and how to pause and for assessing performance.

4. This Dissertation

Figure 2 presents an overview of the five studies comprising this thesis. To achieve the goal of this thesis (i.e., fostering the understanding of pausing as a professional skill by investigating pause effects theoretically and empirically), Study 1 provides a theoretical framework of the medical pause, while four empirical studies of Study 2 to 5 follow based on this framework. After Study 2 validates measures and the task environment of a 2D CBS (i.e., AbcdeSIM [73]), Study 3 and 4 investigate the two processes during pauses (i.e., relaxation and reflection) respectively. Finally, Study 5 tests the generalizability of the findings in another type of CBS environment, Virtual Reality (VR). This structure is reflected in the following chapters.

Chapter 2 (Study 1) explicate the importance, processes, and training of the medical pause. I describe the advantages of pausing for performance and learning (importance), analyze constituent cognitive processes of pausing (processes), and suggest several strategies to train pausing skills in educational programs (training). By introducing concepts from cognitive load theory and educational psychology, I describe how the two major cognitive processes during pauses (i.e., relaxation and reflection) may affect performance and learning.

Figure 2. Overview of the five studies reported in this thesis.



Chapter 3 (Study 2) examines whether the measures and task environments I use have sufficient discriminatory power. Based on theories of complex learning (e.g., 4C/ID [27]), I derive four constructs that represent competency levels: systematicity, accuracy, speed, and cognitive load. These constructs, then, are operationalized to specific measures by using game logs and eye-tracking. These measures are validated as the indicators of the competency levels in a computer-based simulation game for emergency medicine, AbcdeSIM.[73]

Chapter 4 (Study 3) investigates how relaxation pauses affect performance and cognitive load in intense situations. On the assumption that allowing pauses and actually taking pauses are two different constructs, their effects are tested respectively. Pupillometry, an objective and continuous measure of cognitive load, is used to see how participants react to changing situations in the dynamic task environment of AbcdeSIM. The relationship between cognitive load, relaxation during pauses, and performance, is discussed.

Chapter 5 (Study 4) tests the effects of reflective pauses on performance, given instructional support for the reflection provided. By using concepts from complex learning, I propose how to design cognitive and metacognitive aids (CMAs) to support reflection processes during pausing. Assuming reflective pauses with CMAs help to optimize cognitive load and allow for restructuring of mental models, I examine their effects on four aspects of performance and learning: cognitive load, domain-specific performance, domain-general performance, and the structure of cognitive schemas.

Chapter 6 (Study 5) examines whether the indicators found from the previous studies can be applied to VR environments. Although pupillometry is well-known as a reliable technique to measure cognitive load in 2D environments, its applicability to VR environments has not been validated yet. Specifically, the VR display causes light reflexes that confound task-evoked pupillary responses (TEPRs). I validate whether task difficulty can predict cognitive load by using TEPRs corrected for this confounding factor, and whether these TEPRs correlate with performance.

Finally, Chapter 7 (General Discussion) brings the findings from all studies together. By connecting these findings, it discusses how the studies in this thesis contributed to the investigation of the pause effects and deepened the understanding of the medical pause. This project's theoretical and methodological contribution as well as limitations are illustrated to guide future researchers to unfold the research on the medical pause. Also, practical implications for educators and CBS designers are discussed.

Reference

1. Institute of Medicine. *To Err is Human: Building a Safer Health System*. Vol 6. Washington, DC: National Academies Press; 2000.
2. Bates DW, Singh H. Two decades since to err is human: an assessment of progress and emerging priorities in patient safety. *Health Affairs*. 2018;37(11):1736-1743. 10.1377/hlthaff.2018.0738
3. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: The National Academies Press; 2001.
4. Joint Commission on Accreditation of Healthcare Organizations. *Comprehensive Accreditation Manual: CAMH for Hospitals: the Official Handbook*. Oakbrook Terrace, IL: Joint Commission on Accreditation of Healthcare Organizations; 2009.
5. Kelly JJ, Farley H, O'Cain C, et al. A survey of the use of time-out protocols in emergency medicine. *The Joint Commission Journal on Quality and Patient Safety*. 2011;37(6):285-288. 10.1016/S1553-7250(11)37036-5
6. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine*. 2009;360(5):491-499. 10.1056/NEJMsa0810119
7. Altpeter T, Luckhardt K, Lewis JN, Harken AH, Polk Jr HC. Expanded surgical time out: a key to real-time data collection and quality improvement. *J Am Coll Surg*. 2007;204(4):527-532. 10.1016/j.jamcollsurg.2007.01.009
8. Rall M, Glavin R, Flin R. The '10-seconds-for-10-minutes principle'. *Bull R Coll Anaesth*. 2008; 51:2,614-612,616.
9. Mahajan RP. The WHO surgical checklist. *Elsevier BV*. 2011; 25(2): 161-168. 10.1016/j.bpa.2011.02.002
10. Schmutz JB, Eppich WJ. Promoting Learning and Patient Care Through Shared Reflection. *Acad Med*. 2017;92(11):1555-1563. 10.1097/acm.0000000000001688
11. Clapper TC, Leighton K. Incorporating the reflective pause in simulation: A practical guide. *The Journal of Continuing Education in Nursing*. 2020;51(1):32-38. 10.3928/00220124-20191217-07
12. Moulton CA, Epstein RM. Self-monitoring in surgical practice: Slowing down when you should. In: Fry H. KR, ed. *Surgical Education*. Springer, Dordrecht; 2011:169-182.
13. St-Martin L, Patel P, Gallinger J, Moulton CA. Teaching the slowing-down moments of operative judgment. *Surgical Clinics*. 2012;92(1):125-135. 10.1016/j.suc.2011.12.001
14. Melnyk BM. Culture Eats Strategy Every Time: What Works in Building and Sustaining an Evidence-Based Practice Culture in Healthcare Systems. *Worldviews Evid Based Nurs*. 2016;13(2): 99-101. 10.1111/wvn.12161
15. Lee JY, Szulewski A, Young JQ, Donkers J, Jarodzka H, Van Merriënboer JJG. The Medical Pause: Importance, Processes, and Training. *Med Educ*. 2021. 10.1111/medu.14529
16. Sparks EA, Wehbe-Janek H, Johnson RL, Smythe WR, Papaconstantinou HT. Surgical Safety Checklist Compliance: A Job Done Poorly! *Elsevier BV*. 2013;217(5):867-873.e863. 10.1016/j.jamcollsurg.2013.07.393

17. Rydenfalt C, Johansson G, Odenrick P, Akerman K, Larsson PA. Compliance with the WHO Surgical Safety Checklist: deviations and possible improvements. *Int J Qual Health Care.* 2013; 25(2):182-187. 10.1093/intqhc/mzt004
18. Schmutz JB, Kolbe M, Eppich WJ. Twelve tips for integrating team reflexivity into your simulation-based team training. *Med Teach.* 2018;40(7):721-727. 10.1080/0142159x.2018.1464135
19. Trowbridge RL. Twelve tips for teaching avoidance of diagnostic errors. *Med Teach.* 2008;30(5): 496-500. 10.1080/01421590801965137
20. Lee JY, Donkers J, Jarodzka H, Sellenraad G, Van Merriënboer JJG. Different effects of pausing on cognitive load in a medical simulation game. *Comput Human Behav.* 2020; 110: 106385. 10.1016/j.chb.2020.106385
21. Hasler BS, Kersten B, Sweller J. Learner control, cognitive load and instructional animation. *Appl Cogn Psychol.* 2007;21(6):713-729. 10.1002/acp.1345
22. Moulton CA, Regehr G, Mylopoulos M, Macrae HM. Slowing Down When You Should: A New Model of Expert Judgment. *Acad Med.* 2007; 82(Suppl): S109-S116. 10.1097/acm.0b013e3181405a76
23. Schmutz JB, Lei Z, Eppich WJ, Manser T. Reflection in the heat of the moment: The role of in-action team reflexivity in health care emergency teams. *Journal of Organizational Behavior.* 2018;39(6): 749-765. 10.1002/job.2299
24. Sweller J, Van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev.* 2019;31(2):261-292. 10.1007/s10648-019-09465-5
25. Szulewski A, Howes D, Van Merriënboer JJG, Sweller J. From Theory to Practice: The Application of Cognitive Load Theory to the Practice of Medicine. *Acad Med.* 2020;Publish Ahead of Print. 10.1097/acm.0000000000003524
26. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001
27. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
28. Butler DL, Winne PH. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research.* 1995;65(3):245-281. 10.3102/00346543065003245
29. Ishak AW, Ballard DI. Time to Re-Group: A Typology and Nested Phase Model for Action Teams. *Small Group Research.* 2012;43(1):3-29. 10.1177/1046496411425250
30. Bloom BS. *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill Book Company; 1971.
31. Backster A, Teo A, Swift M, Polk HC, Harken AH. Transforming the Surgical "Time-Out" Into a Comprehensive "Preparatory Pause". *J Card Surg.* 2007;22(5):410-416. 10.1111/j.1540-8191.2007.00435.x
32. Shahid S, Thomas S. Situation, Background, Assessment, Recommendation (SBAR) Communication Tool for Handoff in Health Care – A Narrative Review. *Safety in Health.* 2018;4(1). 10.1186/s40886-018-0073-1
33. Rosenbluth G, Destino LA, Starmer AJ, et al. I-PASS Handoff Program: Use of a campaign to effect transformational change. *Pediatric quality & safety.* 2018;3(4):e088.

34. Pronovost PJ, Vohr E. *Safe patients, smart hospitals: how one doctor's checklist can help us change health care from the inside out.* New York, NY: Plume; 2011.
35. Hales BM, Pronovost PJ. The checklist—a tool for error management and performance improvement. *J Crit Care.* 2006;21(3):231-235. 10.1016/j.jcrc.2006.06.002
36. Pronovost PJ, Bo-Linn GW. Preventing Patient Harms Through Systems of Care. *JAMA.* 2012; 308(8):769. 10.1001/jama.2012.9537
37. Thomassen Ø, Storesund A, Søfteland E, Brattebø G. The effects of safety checklists in medicine: a systematic review. *Acta Anaesthesiol Scand.* 2014;58(1):5-18. 10.1111/aas.12207
38. Sibbald M, De Bruin AB, Van Merrienboer JJ. Checklists improve experts' diagnostic decisions. *Med Educ.* 2013;47(3):301-308. 10.1111/medu.12080
39. Moulton CA, Regehr G, Lingard L, Merritt C, Macrae H. 'Slowing Down When You Should': Initiators and Influences of the Transition from the Routine to the Effortful. *J Gastrointest Surg.* 2010; 14(6):1019-1026. 10.1007/s11605-010-1178-y
40. Moulton CA, Regehr G, Lingard L, Merritt C, MacRae H. Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. *Acad Med.* 2010;85(10):1571-1577. 10.1097/ACM.0b013e3181f073dd
41. Huang T-T. *The effects of types of reflective scaffolding and language proficiency on the acquisition of physics knowledge in a game-based learning environment*, New York University; 2016.
42. Nelson BC. Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *Journal of Science Education and Technology.* 2007;16(1):83-97.
43. Ke F. Computer games application within alternative classroom goal structures: cognitive, metacognitive, and affective evaluation. *Educational Technology Research and Development.* 2008; 56(5-6):539-556. 10.1007/s11423-008-9086-5
44. Viberg O, Khalil M, Baars M. Self-regulated learning and learning analytics in online learning environments: A review of empirical research. Paper presented at: Proceedings of the tenth international conference on learning analytics & knowledge2020.
45. Schwan S, Riempp R. The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and Instruction.* 2004;14(3):293-305. 10.1016/j.learninstruc.2004.06.005
46. Tabbers HK, de Koeijer B. Learner control in animated multimedia instructions. *Instructional Science.* 2010;38(5):441-453. 10.1007/s11251-009-9119-4
47. Lowe R. Interrogation of a dynamic visualization during learning. *Learning and Instruction.* 2004; 14(3):257-274. 10.1016/j.learninstruc.2004.06.003
48. Lowe R. Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education.* 1999;14(2):225-244. 10.1007/Bf03172967
49. Schnotz W, Bockheler J, Grzondziel H. Individual and co-operative learning with interactive animated pictures. *European Journal of Psychology of Education.* 1999;14(2):245-265. Doi 10.1007/Bf03172968
50. Fiorella L, Mayer RE. Paper-based aids for learning with a computer-based game. *J Educ Psychol.* 2012;104(4):1074.
51. Ter Vrugte J, De Jong T, Wouters P, Vandercruysse S, Elen J, Van Oostendorp H. When a game supports prevocational math education but integrated reflection does not. *J Comput Assisted Learn.* 2015;31(5):462-480. 10.1111/jcal.12104

52. Kori K, Pedaste M, Leijen Ä, Mäeots M. Supporting reflection in technology-enhanced learning. *Educational Research Review.* 2014;11:45-55. 10.1016/j.edurev.2013.11.003

53. Taub M, Azevedo R, Bradbury AE, Mudrick NV. Self-regulation and reflection during game-based learning. *Handbook of Game-Based Learning.* 2020:239-262.

54. Yang X, Liu Y. Supporting Students' Reflection in Game-Based Science Learning: A Literature Review. In: Li R. CSKS, Iwasaki C., Kwok LF., Kageto M. , ed. *Lecture Notes in Computer Science.* Springer International Publishing; 2021:119-131.

55. Azevedo R, Mudrick NV, Taub M, Bradbury AE. Self-regulation in computer-assisted learning systems. 2019.

56. Azevedo R, Taub M, Mudrick NV, Martin S, Grafsgaard J. Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. *Handbook of self-regulation of learning and performance.* 2018;2.

57. Kalyuga S. Cognitive load theory: How many types of load does it really need? *Educ Psychol Rev.* 2011;23(1):1-19.

58. Schnotz W, Kürschner C. A reconsideration of cognitive load theory. *Educ Psychol Rev.* 2007; 19(4):469-508.

59. Endsley MR. Situation awareness. *Handbook of human factors and ergonomics.* 2006;3:528-542.

60. Tien G, Atkins MS, Zheng B, Swindells C. Measuring situation awareness of surgeons in laparoscopic training. Paper presented at: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications; March, 2010.

61. LeBlanc VR. The effects of acute stress on performance: implications for health professions education. *Acad Med.* 2009;84(10):S25-S33.

62. Yerkes RM, Dodson JD. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology.* 1908;18(5):459-482. 10.1002/cne.920180503

63. Chen O, Castro-Alonso JC, Paas F, Sweller J. Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect. *Educ Psychol Rev.* 2018; 30(2):483-501. 10.1007/s10648-017-9426-2

64. Szulewski A, Roth N, Howes D. 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. *Acad Med.* 2015;90(7):981-987. 10.1097/Acm.0000000000000677

65. McBee E, Ratcliffe T, Picho K, et al. Contextual factors and clinical reasoning: differences in diagnostic and therapeutic reasoning in board certified versus resident physicians. *BMC Med Educ.* 2017;17(1):211.

66. Graafland M, Schraagen JM, Schijven MP. Systematic review of serious games for medical education and surgical skills training. *Br J Surg.* 2012;99(10):1322-1330. 10.1002/bjs.8819

67. Dankbaar M. Serious games and blended learning; effects on performance and motivation in medical education. *Perspect Med Educ.* 2017;6(1):58-60. 10.1007/s40037-016-0320-2

68. Pottle J. Virtual reality and the transformation of medical education. *Future Healthcare Journal.* 2019;6(3):181-185. 10.7861/fhj.2019-0036

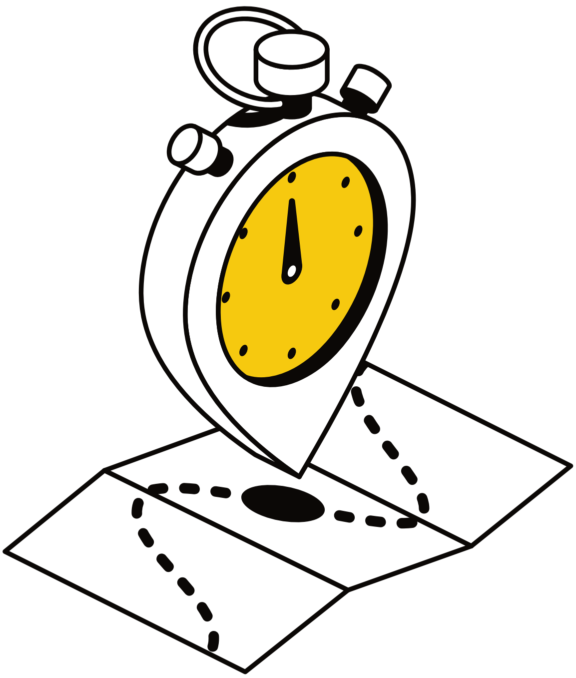
69. Holmqvist K, Andersson R. *Eye-tracking: A comprehensive guide to methods, paradigms and measures.* Lund, Sweden: Lund Eye-Tracking Research Institute; 2017.

70. Tien G, Zheng B, Atkins MS. Quantifying surgeons' vigilance during laparoscopic operations using eyegaze tracking. Paper presented at: MMVR2011.
71. Law B, Atkins MS, Kirkpatrick AE, Lomax AJ. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. Paper presented at: Proceedings of the 2004 symposium on Eye tracking research & applications2004.
72. Szulewski A, Gegenfurtner A, Howes DW, Sivilotti ML, van Merriënboer JJ. Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education.* 2017;22(4):951-968. 10.1007/s10459-016-9725-2
73. *AbcdeSIM* [computer program]. Rotterdam: VirtualMedSchool; 2012.

Chapter 2

The Medical Pause: Importance, Processes, and Training

This chapter is published as:
Lee JY, Szulewski A, Young JQ, Donkers J, Jarodzka H, Van Merriënboer JJG. The medical pause: Importance, processes, and training. Med Educ. 2021. 10.1111/medu.14529



Abstract

Research has shown that taking "timeouts" in medical practice improves performance and patient safety. However, the benefits of taking timeouts, or pausing, is not sufficiently acknowledged in workplaces and training programs. To promote this acknowledgment, we suggest a systematic conceptualization of the medical pause, focusing on its importance, processes, and implementation in training programs. By employing insights from educational and cognitive psychology, we first identified pausing as an important skill to interrupt negative momentum and bolster learning. Subsequently, we categorized constituent cognitive processes for pausing skills into two phases: the decision-making phase (determining when and how to take pauses) and the executive phase (applying relaxation or reflection during pauses). We present a model that describes how relaxation and reflection during pauses can optimize cognitive load in performance. Several strategies to implement pause training in medical curricula are proposed: intertwining pause training with training of primary skills, providing second-order scaffolding through shared control, and employing auxiliary tools such as computer-based simulations with a pause function.

Keywords: Timeout, patient safety, cognitive load, skill development, medical training

1. Introduction

Research has revealed that medical errors are among the leading causes of preventable deaths each year.[1,2] The widespread agreement that healthcare delivery can harm patients has spurred a movement toward quality improvement and diverse strategy application to improve patient safety.[3] The *timeout*, where the current course of action is paused and the parameters of safe care are reassessed, has been adopted as safety strategy in many settings.[4,5] Although studies have shown that taking timeouts significantly decreases patient morbidity and mortality,[6-8] compliance with timeout protocols appears to be low, diluting its clinical benefit.[5,9,10]

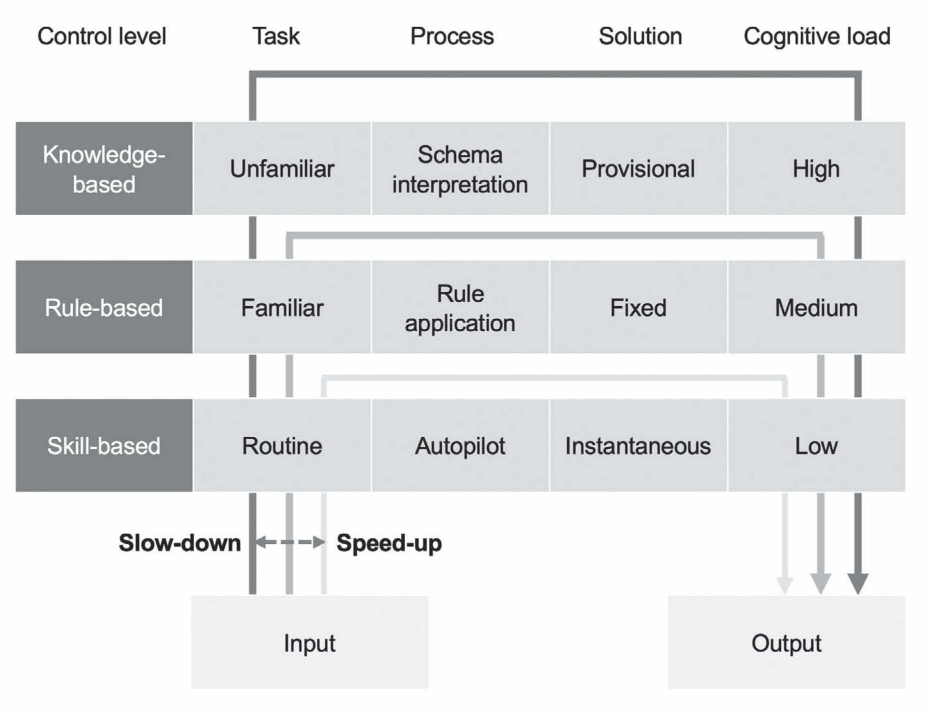
We posit that the basic nature of the timeout is *pausing*, a conscious decision to stop current performance for a physical time that allows for additional cognitive activities. While medical timeouts in general refer to a formal practice with structured protocols in a team setting (e.g., surgical timeout, diagnostic timeout),[11,12] pausing can occur either individually or collectively, with or without formal protocols. Pausing as a general construct and strategy can potentially advance safety and learning in medicine. Yet, to date, we do not have a theoretical understanding of what constitutes a pause and how pausing might benefit clinical outcomes and spur cultural changes.[13]

In the following, we aim to explicate the importance, processes, and training of the medical pause. First, we will focus on the unique advantages of pausing: intercepting negative momentum and promoting learning (Importance). Second, we will describe a new conceptual framework of pausing by analyzing its constituent cognitive processes (Processes). Third, moving from a theoretical basis to practical implementation, we will suggest several strategies to train pausing within the medical curriculum (Training). Based on our combined experience in both educational psychology and clinical work, we will integrate existing theories from various fields (e.g., cognitive psychology and complex learning), resulting in a systematic conceptualization of the medical pause.

2. Importance: The Strengths of Pausing

Problem solving is one of the major approaches to clinical reasoning.[14] Using an elaborated version of dual-process theories[15,16] and cognitive load theory,[17] we will describe how clinical problem solving is guided by three different levels of cognitive control: skill-based, rule-based, and knowledge-based control (Figure 3).

Figure 3. Three levels of cognitive control in real-life problem solving (skill-based, rule-based, and knowledge-based)



Note. The two-way arrow on the right side represents two different directions in focus changing across these levels (slowing down and speeding up)

Skill-based control (SC) is activated when dealing with a routine task, using cognitive rules internalized through repetitive practice. When faced with a patient who has ventricular fibrillation, an experienced emergency physician can quickly ensure that the team prioritizes chest compressions and defibrillation. With years of training and experience, this process becomes automatic and intuitive (i.e., autopilot). When problems are not entirely routine but familiar enough to apply internal or external rules, *rule-based control* (RC) is initiated.[16] Using checklists during surgical timeouts or following step-by-step instructions for procedures like central line insertion are examples of this process. For atypical and novel problems, *knowledge-based control* (KC) is stimulated, where "knowledge" means mental models or schemas of a domain that reflect how concepts in the domain are structured.[18] When diagnosing a patient with a complex presentation (e.g., adrenal insufficiency as a cause of hypotension), a physician may need to consider alternative diagnoses by reflecting on similar cases and building upon relevant knowledge. Whereas the level of conscious control and its derived cognitive load are highest in the case of KC, they are lowest when practicing SC.[19]

In real-life problem solving, these three levels run in parallel, with weighted focus depending on the activity.[16] Figure 3 depicts how this focus shifts across the levels: *slowing down*[20] as the focus moves from SC to KC, and *speeding up* when the opposite is true. During a routine operation, a surgeon who notices atypical anatomy would pause the procedure to reassess a list of alternatives and carefully apply the selected alternative (i.e., slowing down). Once this step is completed, this surgeon would return to the faster autopilot mode (i.e., speeding up). Note that this transition is feasible only for experienced performers since novices can neither identify cues that initiate the transition nor effectively change their focus as they lack structured schemas and skills.

While the speeding-up process has been highlighted in the professional skill acquisition literature,[21] the slowing-down process has been emphasized in the context of safety culture and expertise maintenance in medicine.[22] Moulton et al. identified slowing down as a crucial component of medical expertise, coining the phrase "slowing down when you should."[23] They characterized pausing as the most extreme form of slowing down.[24] While slowing down has been described as a cognitive process that may accompany physical manifestation, we further specify pausing as a cognitive and temporal process grounded in the physical world. This physicality grants pausing two unique advantages: pausing allows for (1) explicit interception of negative momentum, and (2) dedicated time for necessary processing at the KC level, such as learning.

In experimental psychology, *negative momentum* is a state stimulated to move away from desired goals.[25] Using an analogy from physical laws, the psychological state keeps its current motion, unless extra force is applied to change that state.[26] We identified two examples of negative momentum in medicine: the *hurry-up syndrome*[27] and *drifting*.[24] The term hurry-up syndrome has been used in aviation safety research which reveals that time pressure is significantly correlated with errors and safety risks.[27] In medicine, time pressure and clinical acuity can cause the hurry-up syndrome, resulting in medical errors and safety issues. Facing this negative momentum, novices tend to act quickly but inappropriately. To encourage them to rather stay undetermined and take a pause, experienced surgeons teach their residents the phrase, "Do not just do something, stand there."[8]

While this undetermined neutral status can be beneficial sometimes, mind wandering or inattentiveness can be less advantageous, comprising another example of negative momentum.[28] Contrary to the hurry-up syndrome, drifting is the distracted state arising from "boring" tasks and complacency.[24] During routine tasks in an operation, surgeons could become careless and join extraneous chatting with colleagues, failing to monitor adverse events as they emerge.[24] They inappropriately remain at the SC level, without allocating the cognitive resources that are freed up by automatic processing to essential activities.

After direct interruption of these types of negative momentum, pausing inserts a certain period of time that not only makes the interruption stronger but also facilitates additional processing at the KC level. We postulate that this processing promotes learning since it allows for a reconstruction of schemas. Ericsson[29] argued that expert performance can be maintained by continuously seeking out learning opportunities, referred to as *deliberate practice*. Deliberate practice includes resisting the tendency toward complacency and deliberately reconstructing schemas by spending additional time analyzing everyday performance.[29] Since learning opportunities from pausing arise during performance, they are strongly fostered by fresh memories and embodied experiences that pre- or post- timeouts cannot provide.

The benefits of pausing are evident and long-term compared to the costs. Rall et al.[8] have shown that, if trained sufficiently, only 10 seconds of timeout can be effective to improve performance in an emergency setting. The time lost taking timeouts is offset by improved team action, while the delay does not significantly jeopardize performance.[8] Moreover, enhanced safety will not only prevent patients from suffering from complications but also protect healthcare professionals from suffering psychologically from medical errors.[30]

3. Processes: Key Components of Pausing

We identified two phases of pausing: the decision-making phase to determine when and how to take pauses, and the executive phase to apply cognitive activities such as relaxation and reflection during the pauses. To make pausing skills effective, multiple cognitive and metacognitive processes must be functioning across these phases.

3.1. Decision-Making Phase

While negative momentum should be interrupted, there is also *positive momentum* that should not be interrupted but rather supported during medical workflow; in other words, “not slowing down when you should not.” In extreme circumstances where expediency must be prioritized (e.g., emergency cesarean section), the opportunity to slow down does not exist. In team performance, taking an individual timeout could interrupt collective performance. Literature in psychology and brain sciences has confirmed that slowing down may impair performance depending on the situation.[31,32] Therefore, deciding on whether or not to pause becomes a highly advanced clinical judgment, involving time management to assess the time available and the risk of using the time.[8]

Planning. The deployment of pauses can take two forms: proactive planning or responsive improvisation.[33] Pausing can be proactively planned before a task, by identifying landmines that require special attention.[33] These landmines are determined by assessing the task complexity and the performer's own capability. Through this assessment and mental simulation, one can establish a game plan[33] where measures for the landmines are arranged. For instance, a surgeon who previously experienced difficulties performing lung resection might plan a pause before the challenging portion of the procedure. Expert performers are good game planners who excel in predicting what is going to happen during a procedure, thanks to their superior mental simulation.[33]

Improvising. Responsively improvised pauses, on the other hand, take place when encountering unexpected events (e.g., recognizing an abnormality in patients' anatomy during surgery) or unsolvable problems (e.g., initial treatment not working during resuscitation). Realizing unexpected landmines along the way, performers must make emergency decisions on taking a pause. This form of decision-making is extra demanding because it involves the adjustment of the original game plan, quick problem solving, accurate calculation of the time available, and constant vigilance for signals to pause.[33]

Self-monitoring. Although this vigilance is a prerequisite for pause deployment, the negative momentum is sometimes the result of vigilance decrement.[34] In considering this dilemma, we argue that introducing metacognition[35], another level of cognitive processes that can oversee cognition itself, might be a solution. Self-monitoring is a metacognitive process to supervise one's own thoughts and mental status.[18] Overseeing one's own monitoring capability could be a measure to help prevent vigilance decrement.

Cognitive load. We use cognitive load theory (CLT)[19,36] to further explain *what* and *how* to self-monitor. CLT presumes that diverse mental processes evoke cognitive load on working memory and that task performance deteriorates if the load exceeds working memory capacity. Research has shown that vigilance produces a tangible amount of cognitive load,[34,37] which means performers should always reserve sufficient capacity for this process in working memory.[23] To stress the importance of this reservation, we suggest a new typology of cognitive load: primary load (PL) caused by domain-specific primary tasks (e.g., a surgical task), secondary load (SL) from domain-general processes that support the primary tasks (e.g., self-monitoring during surgery), and extraneous load (EL) that does not contribute to the tasks (e.g., distracting conversations during surgery). Thus, the targets of self-monitoring are narrowed down to two constructs: whether the room for SL is reserved and whether the total load is below the limit.

Conceptualizing cognitive load and its typology with explicit terminology guides performers on how to self-monitor. It facilitates the conscious control of behavior,[38] providing a useful internal cue for self-monitoring that can be measured by introspection.[39] In various medical domains, cognitive load is known as a valid indicator of expertise, performance, and learning.[39-42] The internality of the cue is valuable in many domains of medicine, where access to information on decision support may be limited and professionals often work alone or with less experienced staff who may not be able to provide high-quality feedback.[23]

Collective decision-making. In team performance, more external cues and interventions are necessary. To develop external cues, the team could use real-time clinical decision support[43] and encourage themselves to think aloud during performance. Based on this shared mental profile, the team leader or a separate overseer could make decisions to improvise pauses. When the cues are insufficient, forced pauses initiated by formal timeout protocols can be of benefit.[44] I-PASS[45] and SBAR[46] are good examples of organizational efforts to implement interventions to facilitate medical pauses.

3.2. Executive Phase

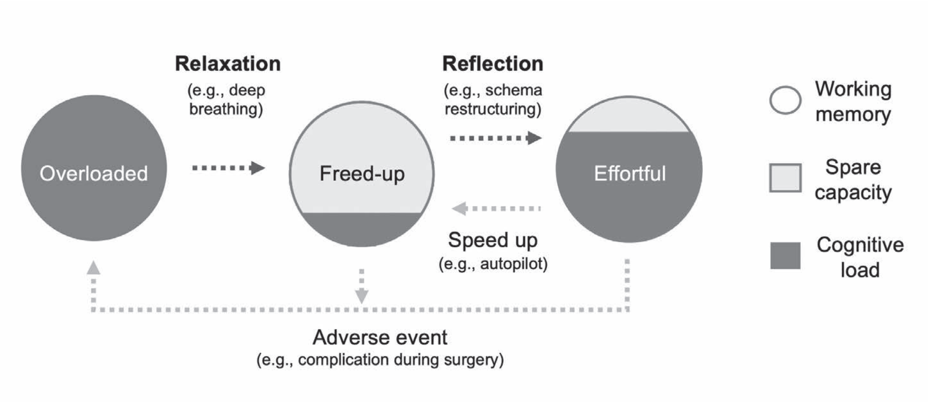
Once a decision to pause is made, the processes to optimize internal and external resources ensue to redirect the negative momentum. In the stress literature, this optimization is essential to titrate the stress level below distress while maintaining the eustress status, thereby maximizing performance.[47,48] Similarly in CLT, to promote performance and learning, cognitive load should be kept below working memory capacity, but not too low in order to maintain PL and SL.

We identified two types of processes to find the optimal level: *relaxation* that reduces overload in working memory, and *reflection* that interconnects working memory and long-term memory to restructure schemas. In difficult resuscitation cases, a physician may pause the case to calm down (relaxation) and "recap" the current situation by summarizing previous performance, asking for others' ideas, and preparing next priorities, based on experience and knowledge (reflection). Experienced physicians are reportedly better skilled in these coping strategies,[41] likely because of their honed reasoning based on the assessment of internal and external resources in given situations.[49]

Relaxation. Figure 4 describes the state transition of the cognitive load level triggered by several factors as well as relaxation and reflection. The state of being *overloaded*, sometimes referred to as "helmet fire," is caused by adverse events such as time pressure, life-threatening complications, and overwhelmingly complex task environments. This results in acute stress and heightened emotions, contributing to either PL or EL depending on the task.[50] Through relaxation processes, the overloaded state can transition to the *freed-up* state, as the source of overload is managed or the depletion of working memory is recovered through cognitive rest.[51]

Among several relaxation techniques,[52] deep breathing has been recommended for clinical settings as the fastest and simplest technique.[53] Deep breathing, or diaphragmatic breathing, works as a brief form of meditation inserted within performance.[54] Empirical studies show that it reduces stress, anxiety, self-doubt, and cortisol levels,[55] also benefiting performance and learning, such as working memory enhancement, and motor skill acquisition and retention.[56,57]

Figure 4. The cognitive load state transition



Note. The transition is triggered by relaxation and reflection during pauses, as well as speeding up and adverse events. Since some tasks do not allow for the opportunity to take a pause, deciding on whether or not to pause becomes a highly advanced clinical judgement.

Reflection. The freed-up state is a transient phase where relaxation or the speeding up processes have created spare capacity for potentially any type of cognitive load to be imposed. This spare capacity can be filled with either positive sources (PL and SL) or rather negative sources (EL). Through reflection, positive load is activated and the freed-up state transitions to the *effortful* state. In this state, cognitive resources are reallocated more efficiently to maintain control over performance.

Reflection is a form of cognitive restructuring vital in every practice of medicine.[58] By interconnecting working memory and long-term memory, it allows for creative solutions or a "fresh look" by redefining the given problem. Reflection may improve learning as well as performance,[59] since the act of cognitive restructuring promotes the development of mental models. Diverse medical training has included reflective activities to promote learning processes.[12,60]

Drifting. When spare capacity is devoted to EL (e.g., listening to the radio during surgery), the freed-up state transitions to rather negative states such as drifting, where cognitive resources are not invested in essential monitoring activities. By removing the source of EL (e.g., turning off the radio), this state may revert to the freed-up state and move forward to the effortful state by increasing SL.

4. Training: Practical Implementation

Now that we understand the importance and processes of pausing, the next issue to be addressed how to train it. We suggest that (1) pausing should be trained within the regular medical curriculum from the start (intertwining), (2) customized support should be provided until the learner can make pauses independently (scaffolding), and (3) auxiliary tools (e.g., checklists) and learning environments (e.g., computer-based simulation) can be used.

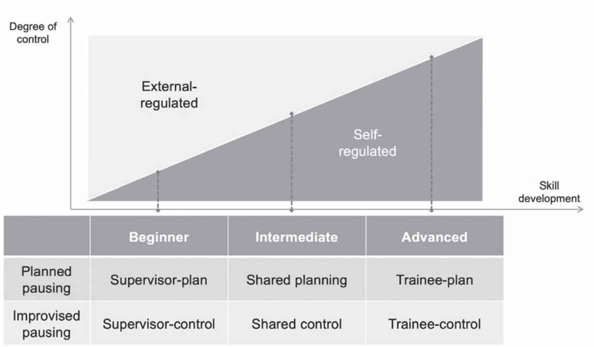
4.1. Intertwining and Scaffolding

Pausing skills cannot be taught by a separate course alone since it is a domain-general secondary skill that is interlaced with domain-specific primary tasks.[18] Research in education has shown that teaching secondary skills outside of primary skills training consistently fails.[61] Moreover, transitioning between different cognitive control levels is not feasible without domain-specific schemas in long-term memory. Thus, it is necessary to *intertwine* pausing skills training with a primary skills training curriculum from a very early stage.
Given this necessity, the question that arises next is how to guide novices throughout the curricula. Making a pause meaningful is a highly advanced skill that requires clinical expertise and workplace experience. According to instructional design models,[18] training programs for complex skills should be presented in a simple-to-complex manner, through scaffolding where support and guidance gradually decrease as learners become more experienced. While scaffolding is generally referred to as a technique for domain-specific skills, *second-order scaffolding* is the technique designated for domain-general skills.[18]

Figure 5 demonstrates a second-order scaffolding model for pausing skills. Since practical skills such as pausing can only be taught by hands-on experience,[62] some degree of direct control to initiate pauses should be granted to learners throughout a program. During this program, supervisors have a dual responsibility: they must transfer short-term and direct control to learners while maintaining global control of the training.[63] To find a balance, supervisors should first assess the learner's level of competence, after which both supervisor and resident can negotiate the distribution of control. The degree of supervisor control should be maximized for beginners (i.e., external-regulated) and gradually reduces in favor of learner control as the learner becomes experienced (i.e., self-regulated).

To teach proactive planning of pausing, learners should have the opportunity to plan for a scenario, rather than merely having a brief look at the patient chart. During this activity, supervisors should provide a framework to predict the landmines that require special measures. By explicitly locating the required focus, learners can deploy either informal or formal timeouts. To maximize learning opportunities, this plan should be thorough and specific to a particular case (e.g., "for this scenario, I will pause to be extra reflective at this point").[63]

Figure 5. Second-order scaffolding to improve pausing skills.



Note. While first-order scaffolding applies to domain-specific skills, second-order scaffolding pertains to domain-general skills. As a learner becomes more experienced, the supervisor's support in planning

Learners can train improvisation of pausing by modeling the initiation of in-action pauses. In the case of novices, the supervisor should take full control of such initiation, while the learner acts as observer. By attending various unexpected real-life events (e.g., manpower shortage, malfunctioning equipment), the learner can develop his or her own measures (e.g., "if this happens to me, I will quickly assess how much time I have and initiate a timeout"). As the learner becomes more experienced, the supervisor's support in planning and improvising would gradually disappear, eventually resulting in full control by the learner.

In many training programs, prompting learners to construct such If-Then cognitive rules[18] as an anticipatory strategy is already a common teaching method. Thus, integrating these cognitive rules for pausing might be easily embedded in existing programs. Additionally, using auxiliary tools such as checklists or debriefings can be a beneficial approach to scaffolding. For instance, diagnostic checklists can help learners to train what to reflect on during pauses. Debriefing about how the executed pauses contributed to performance could also afford additional learning opportunities.

4.2. Simulations With a Pause Function

Simulated task environments offer favorable learning opportunities for medical training.[18] They present diverse scenarios without risking patient safety or loss of materials, provide better control over the format of learning content (e.g., sequence of tasks, support and guidance), and make it possible to pace the task progress. This flexibility in format and pace facilitates second-order scaffolding for pausing skills: Instructors can tailor their level of intervention to teach pausing by modulating the task progress, as learners' competency grows.

A promising way of training pausing skills is to use computer-based simulation (CBS) which allows for easy installment and practice of pause functions. Diverse CBS formats, such as web-based multimedia, serious games, and virtual or augmented reality, have been used to replicate clinical settings.[40,64] In any given format, learners can practice the planning and improvisation of pauses merely by pressing a pause button. Empirical studies have shown that novices experience more cognitive load during CBS training than experts do, and that pausing may significantly help them to manage this load, especially during moments of panic (e.g., when virtual patients have a seizure).[40,65] However, since students reportedly tend not to understand the importance of pausing,[65] CBS training with a pause function should include specific instructions for students to acknowledge the importance to initiate effective pauses by referring to relevant cues.

Another advantage of using CBS is that it produces objective real-time data (e.g., log data of serious games) about performance that can be used as these cues. By integrating it with other types of data such as eye movements, the data can become even richer. Eye-tracking is a novel technique that provides specific information about users' cognitive processes during human-computer interaction.[66] For instance, the level of self-monitoring or vigilance can be measured by recording how gazes are allocated across different objects.[40] The level of cognitive load can be also measured in real time via pupillometry.[41,67]

5. Discussion

In this article, we identified pausing as a professional skill used to interrupt negative momentum and facilitate reflective performance and learning. The constituents of pausing skills were categorized into two phases: the decision-making phase (planning, improvising, and self-monitoring) and the executive phase (relaxation and reflection). We argued that training programs should be systematically designed to cultivate pausing skills.

The main contribution of this work lies in that it connects key concepts in safety culture that have been discussed in different contexts and brings them back to the fundamental principles of the medical pause. We have linked the practical observations and strategies from studies on slowing down phenomena identified by Moulton et al.[20,23,33] and checklists by Pronovost et al.[68-70] We then recognized the cognitive mechanisms shared between the basic concepts in these studies and integrated them within a broad range of clinical activity. This will move the literature one step further as it supports the arguments of the previous studies on formal or informal timeouts and allows future researchers to build upon the literature through our newly developed understanding. Moreover, since the idea of the medical pause can be easily applied to various health professions, more practitioners may consider how to adopt a simple habit of pausing, allowing the existing discussions on slowing down and checklists to develop.

Another contribution of this work is that it provides practical techniques to implement pausing in training programs as well as examples of application, which serves as a translational education effort. This provision may be of value to educators seeking to design better training programs that foster the habit of reflective practice. They can start by introducing simple applications such as checklists that prompt learners to plan and improvise pauses on their own. Moreover, we discussed cognitive constructs using the new terms of primary and secondary cognitive load, which may facilitate explicit communication for supervisors and learners to jointly develop second-order scaffolding.

We identify two limitations to this work. First, we did not consider other factors that might influence the practice of pausing (e.g., personality, emotions, social pressure). For instance, social pressure to make enjoyable conversation during an operation would likely affect the decision-making of pausing. This social aspect of pausing can be further studied in the field of social capital in medicine. Second, since we mainly focused on theoretical conceptualization, empirical evidence is lacking. We, therefore, welcome future studies that empirically verify the effects of medical pausing on performance and cognitive load, and the pragmatic consequences of pause training.

Medical practitioners, being either overloaded or underloaded, can lose control over their performance leading to undesirable consequences. To improve patient safety by effecting cultural and fundamental changes, the acknowledgment of the medical pause should be nurtured through communication, education, and skill-building opportunities. As a result, mastering the medical pause may become an integral part of clinical expertise, in such a way that medical professionals strategically deploy pauses in the heat of the moment by knowing when and how to pause.

References

1. Institute of Medicine. *To Err is Human: Building a Safer Health System.* Vol 6. Washington, DC: National Academies Press; 2000.
2. Bates DW, Singh H. Two decades since to err is human: an assessment of progress and emerging priorities in patient safety. *Health Affairs.* 2018;37(11):1736-1743. 10.1377/hlthaff.2018.0738
3. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: The National Academies Press; 2001.
4. Joint Commission on Accreditation of Healthcare Organizations. *Comprehensive Accreditation Manual: CAMH for Hospitals: the Official Handbook.* Joint Commission on Accreditation of Healthcare Organizations; 2009.
5. Kelly JJ, Farley H, O'Cain C, et al. A survey of the use of time-out protocols in emergency medicine. *The Joint Commission Journal on Quality and Patient Safety.* 2011;37(6):285-288. 10.1016/S1553-7250(11)37036-5
6. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine.* 2009;360(5):491-499. 10.1056/NEJMsa0810119
7. Altpeter T, Luckhardt K, Lewis JN, Harken AH, Polk Jr HC. Expanded surgical time out: a key to real-time data collection and quality improvement. *J Am Coll Surg.* 2007;204(4):527-532. 10.1016/j.jamcollsurg.2007.01.009
8. Rall M, Glavin R, Flin R. The '10-seconds-for-10-minutes principle'. *Bull R Coll Anaesth.* 2008; 51:2,614-612,616.
9. Sparks EA, Wehbe-Janek H, Johnson RL, Smythe WR, Papaconstantinou HT. Surgical Safety Checklist Compliance: A Job Done Poorly! *Elsevier BV.* 2013;217(5):867-873.e863. 10.1016/j.jamcollsurg.2013.07.393
10. Rydenfalt C, Johansson G, Odenrick P, Akerman K, Larsson PA. Compliance with the WHO Surgical Safety Checklist: deviations and possible improvements. *Int J Qual Health Care.* 2013; 25(2):182-187. 10.1093/intqhc/mzt004
11. Mahajan RP. The WHO surgical checklist. *Elsevier BV.* 2011;25(2):161-168. 10.1016/j.bpa.2011.02.002
12. Schmutz JB, Eppich WJ. Promoting Learning and Patient Care Through Shared Reflection. *Acad Med.* 2017;92(11):1555-1563. 10.1097/acm.0000000000001688
13. Melnyk BM. Culture Eats Strategy Every Time: What Works in Building and Sustaining an Evidence-Based Practice Culture in Healthcare Systems. *Worldviews Evid Based Nurs.* 2016;13(2): 99-101. 10.1111/wvn.12161
14. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ.* 2002;324(7339):729-732. 10.1136/bmj.324.7339.729
15. Evans JSBT. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annu Rev Psychol.* 2008;59(1):255-278. 10.1146/annurev.psych.59.103006.093629
16. Olsen SE, Rasmussen J. The reflective expert and the prenovice: notes on skill-, rule-and knowledge-based performance in the setting of instruction and training. In: *Developing skills with information technology.* Wiley; 1989:9-33.

17. Sweller J, Van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev.* 2019;31(2):261-292. 10.1007/s10648-019-09465-5
18. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
19. Van Merrienboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ.* 2010;44(1):85-93. 10.1111/j.1365-2923.2009.03498.x
20. Moulton CA, Regehr G, Mylopoulos M, Macrae HM. Slowing Down When You Should: A New Model of Expert Judgment. *Acad Med.* 2007; 82(Suppl): S109-S116. 10.1097/acm.0b013e3181405a76
21. Dreyfus HL, Dreyfus SE. *Mind over machine.* New York: The Free Press; 1986.
22. Ericsson KA. Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. *Acad Med.* 2004;79(10):S70-S81. 10.1097/00001888-200410001-00022
23. Moulton CA, Epstein RM. Self-monitoring in surgical practice: Slowing down when you should. In: *Surgical Education.* Springer, Dordrecht; 2011:169-182.
24. Moulton CA, Regehr G, Lingard L, Merritt C, MacRae H. Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. *Acad Med.* 2010;85(10):1571-1577. 10.1097/ACM.0b013e3181f073dd
25. Briki W, Markman KD, Coudevylle G, Sinnapah S, Hue O. Momentum sequence and environmental climate influence levels of perceived psychological momentum within a sport competition. 2015:1-8. 10.1080/17461391.2015.1062566
26. Deemer ED, Derosa PA, Duhon SA, Dotterer AM. Psychological Momentum and Inertia: Toward a Model of Academic Motivation. *Journal of Career Development.* 2019:089484531984884. 10.1177/0894845319848847
27. McElhatton J, Drew C. Time pressure as a causal factor in aviation safety incidents- The 'hurry-up' syndrome. Paper presented at: International Symposium on Aviation Psychology, 7 th, Columbus, OH1993.
28. Smallwood J, Mrazek MD, Schooler JW. Medicine for the wandering mind: mind wandering in medical practice. *Med Educ.* 2011;45(11):1072-1080. 10.1111/j.1365-2923.2011.04074.x
29. Ericsson KA. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance.* 2006;38: 685-705.
30. Wallace JE, Lemaire JB, Ghali WA. Physician wellness: a missing quality indicator. *The Lancet.* 2009;374(9702):1714-1721. 10.1016/s0140-6736(09)61424-0
31. Beilock SL, Bertenthal BI, Hoerger M, Carr TH. When does haste make waste? Speed-accuracy tradeoff, skill level, and the tools of the trade. *J Exp Psychol Appl.* 2008;14(4):340.
32. Heitz RP. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front Neurosci.* 2014;8. 10.3389/fnins.2014.00150
33. Moulton CA, Regehr G, Lingard L, Merritt C, Macrae H. 'Slowing Down When You Should': Initiators and Influences of the Transition from the Routine to the Effortful. *J Gastrointest Surg.* 2010; 14(6):1019-1026. 10.1007/s11605-010-1178-y

34. Al S, Tariq, Mir, Alawar, Babiloni, Al N. Vigilance Decrement and Enhancement Techniques: A Review. *Brain Sciences.* 2019;9(8):178. 10.3390/brainsci9080178

35. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001

36. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* 2014; 36(5): 371-384. 10.3109/0142159X.2014.889290

37. Warm JS, Parasuraman R, Matthews G. Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society.* 2008;50(3):433-441. 10.1518/001872008x312152

38. Lingard L, Haber RJ. Teaching and learning communication in medicine: a rhetorical approach. *Academic medicine: Journal of the Association of American Medical Colleges.* 1999;74(5): 507-510. 10.1097/00001888-199905000-00015

39. Blissett S, Sibbald M, Kok E, van Merrienboer J. Optimizing self-regulation of performance: is mental effort a cue? *Advances in Health Sciences Education.* 2018;23(5):891-898. 10.1007/s10459-018-9838-x

40. Lee JY, Donkers J, Jarodzka H, van Merriënboer JJ. How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Comput Human Behav.* 2019;99:268-277. 10.1016/j.chb.2019.05.035

41. Szulewski A, Roth N, Howes D. 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. *Acad Med.* 2015;90(7):981-987. 10.1097/Acm.0000000000000677

42. Young JQ, Van Dijk SM, O'Sullivan PS, Custers EJ, Irby DM, Ten Cate O. Influence of learner knowledge and case complexity on handover accuracy and cognitive load: results from a simulation study. *Med Educ.* 2016;50(9):969-978. 10.1111/medu.13107

43. Shear T, Deshur M, Avram MJ, et al. Procedural Timeout Compliance Is Improved With Real-Time Clinical Decision Support. *Journal of Patient Safety.* 2018; 14(3): 148-152. 10.1097/pts.0000000000000185

44. Ariga A, Lleras A. Brief and rare mental "breaks" keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition.* 2011; 118(3):439-443. 10.1016/j.cognition.2010.12.007

45. Rosenbluth G, Destino LA, Starmer AJ, et al. I-PASS Handoff Program: Use of a campaign to effect transformational change. *Pediatric quality & safety.* 2018;3(4):e088.

46. Shahid S, Thomas S. Situation, Background, Assessment, Recommendation (SBAR) Communication Tool for Handoff in Health Care – A Narrative Review. *Safety in Health.* 2018;4(1). 10.1186/s40886-018-0073-1

47. LeBlanc VR. The effects of acute stress on performance: implications for health professions education. *Acad Med.* 2009;84(10):S25-S33.

48. Yerkes RM, Dodson JD. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology.* 1908;18(5):459-482. 10.1002/cne.920180503

49. McBee E, Ratcliffe T, Picho K, et al. Contextual factors and clinical reasoning: differences in diagnostic and therapeutic reasoning in board certified versus resident physicians. *BMC Med Educ.* 2017;17(1):211.

50. Szulewski A, Howes D, Van Merriënboer JJG, Sweller J. From Theory to Practice: The Application of Cognitive Load Theory to the Practice of Medicine. *Acad Med.* 2020;Publish Ahead of Print. 10.1097/acm.0000000000003524

51. Chen O, Castro-Alonso JC, Paas F, Sweller J. Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect. *Educ Psychol Rev.* 2018; 30(2):483-501. 10.1007/s10648-017-9426-2

52. Parnabas VA, Mahamood Y, Parnabas J, Abdullah NM. The relationship between relaxation techniques and sport performance. *Universal Journal of Psychology.* 2014;2(3):108-112. 10.13189/ujp.2014.020302

53. Young JQ, Wachter RM, ten Cate O, O'Sullivan PS, Irby DM. Advancing the next generation of handover research and practice with cognitive load theory. *BMJ quality & safety.* 2016;25(2): 66-70. 10.1136/bmjqs-2015-004181

54. Gina Paul, Barb Elam, J. Verhulst S. A Longitudinal Study of Students' Perceptions of Using Deep Breathing Meditation to Reduce Testing Stresses. *Teach Learn Med.* 2007;19(3):287-292. 10.1080/10401330701366754

55. Perciavalle V, Blandini M, Fecarotta P, et al. The role of deep breathing on stress. *Neurol Sci.* 2017;38(3):451-458. 10.1007/s10072-016-2790-8

56. Khng KH. A better state-of-mind: deep breathing reduces state anxiety and enhances test performance through regulating test cognitions in children. *Cognition and Emotion.* 2016:1-9. 10.1080/02699931.2016.1233095

57. Yadav G, Mutha PK. Deep Breathing Practice Facilitates Retention of Newly Learned Motor Skills. *Sci Rep.* 2016;6(1):37069. 10.1038/srep37069

58. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. *Med Teach.* 2009; 31(8):685-695. 10.1080/01421590903050374

59. Ohtani K, Hisasaka T. Beyond intelligence: a meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning.* 2018; 13(2):179-212. 10.1007/s11409-018-9183-8

60. Branch Jr WT. Small-group teaching emphasizing reflection can positively influence medical students' values. *Acad Med.* 2001;76(12):1171-1172.

61. Tricot A, Sweller J. Domain-specific knowledge and why teaching generic skills does not work. *Educ Psychol Rev.* 2014;26(2):265-283. 10.1007/s10648-013-9243-1

62. Schön DA. Educating the reflective practitioner: Toward a new design for teaching and learning in the professions. *Australian Journal of Adult Learning.* 2010;50(2):448-451.

63. St-Martin L, Patel P, Gallinger J, Moulton C-a. Teaching the slowing-down moments of operative judgment. *Surgical Clinics.* 2012;92(1):125-135. 10.1016/j.suc.2011.12.001

64. Pantelidis P, Chorti A, Papagiouvanni I, et al. Virtual and Augmented Reality in Medical Education. In: Tsoulfas G, ed. *Medical and Surgical Education: Past, Present and Future.* London, UK: IntechOpen; 2018:77-97.

65. Lee JY, Donkers J, Jarodzka H, Sellenraad G, Van Merriënboer JJG. Different effects of pausing on cognitive load in a medical simulation game. *Comput Human Behav.* 2020; 110: 106385. 10.1016/j.chb.2020.106385

66. Holmqvist K, Andersson R. *Eye-tracking: A comprehensive guide to methods, paradigms and measures.* Lund, Sweden: Lund Eye-Tracking Research Institute; 2017.

67. Szulewski A, Gegenfurtner A, Howes DW, Sivilotti ML, van Merriënboer JJ. Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education.* 2017; 22(4): 951-968. https://link.springer.com/content/pdf/10.1007%2Fs10459-016-9725-2.pdf.

68. Pronovost PJ, Vohr E. *Safe patients, smart hospitals: how one doctor's checklist can help us change health care from the inside out.* Soc Nuclear Med; 2011.

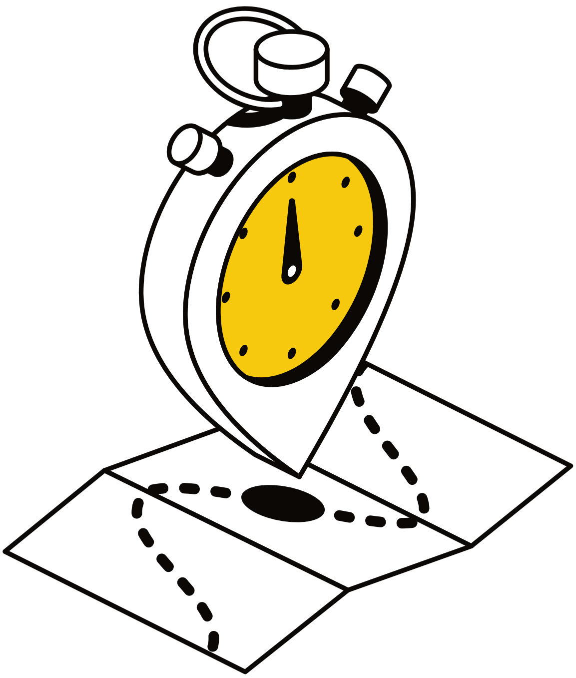
69. Hales BM, Pronovost PJ. The checklist—a tool for error management and performance improvement. *J Crit Care.* 2006;21(3):231-235. 10.1016/j.jcrc.2006.06.002

70. Pronovost PJ, Bo-Linn GW. Preventing Patient Harms Through Systems of Care. *JAMA.* 2012; 308(8):769. 10.1001/jama.2012.9537

Chapter 3

How Prior Knowledge Affects Problem-Solving Performance in a Medical Simulation Game: Using Game-Logs and Eye-Tracking

This chapter is published as:
Lee JY, Donkers J, Jarodzka H, van Merriënboer JJ. How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. Comput Human Behav. 2019;99:268-277. 10.1016/j.chb.2019.05.035



Abstract

Computer-based simulation games provide an environment to train complex problem-solving skills. Yet, it is largely unknown how the in-game performance of learners varies with different levels of prior knowledge. Based on theories of complex skill acquisition (e.g., 4C/ID), we derive four performance aspects that prior knowledge may affect: (1) systematicity in approach, (2) accuracy in visual attention and motor reactions, (3) speed in performance, and (4) cognitive load. This study aims to empirically test whether prior knowledge affects these four aspects of performance in a medical simulation game for resuscitation skills training. Participants were 24 medical professionals (experts, with high prior knowledge) and 22 medical students (novices, with low prior knowledge). After pre-training, they all played one scenario, during which game logs and eye movements were collected. A cognitive-load questionnaire ensued. During game play, experts demonstrated a more systematic approach, higher accuracy in visual selection and motor reaction, and a higher performance speed than novices. Their reported levels of cognitive load were lower. These results indicate that prior knowledge has a substantial impact on performance in simulation games, opening up the possibility of using our measures for performance assessment.

Keywords: Assessment, prior knowledge, simulation game, serious game, cognitive load, eye-tracking

1. Introduction

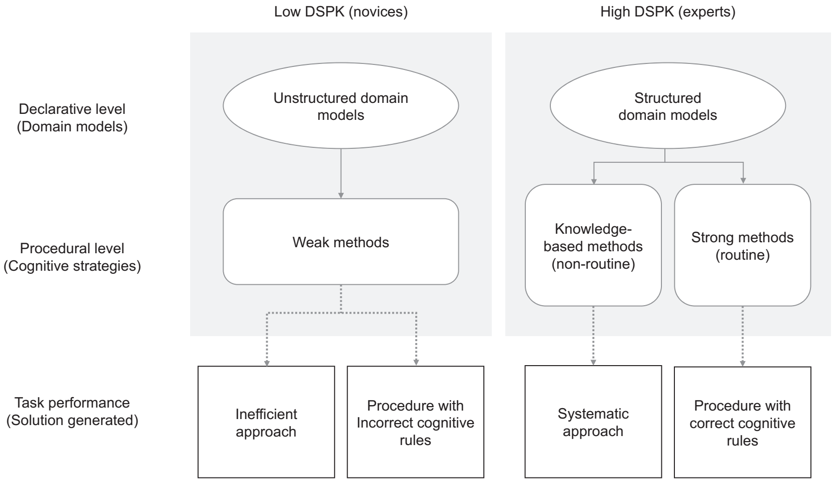
Computer-based simulation games (CBSG) are effective learning environments for complex skills. As simulations, they approximately replicate the complexity of real-life situations,[1] and as computer games, they provide a package of problems that are causally connected, based on learners' interaction with the game.[2] In this simulated problem-solving environment, learners can train specific professional skills in areas such as aviation, business management, and medicine.[3-5]

However, CBSGs face a challenge in that the performance of a learner in the game is difficult to assess via traditional measurements such as achievement tests.[6] This challenge is mainly due to the open-ended nature of CBSGs,[7] which allows for a large number of different behaviors. Therefore, recent research has focused on tracking users' in-game behaviors by looking at game data such as *serious game analytics*.[6,8,9] These studies identified several limitations: Data analysis without involving educational theoretical principles often fails to fully account for students' performance,[6] game logs without translation into high-level meaningful actions can yield confounding information,[10] some important factors such as timing cannot be explained by analyzing sequences of events only,[11] and empirical studies about how game data can be informative for performance assessment are scarce.[6,12]

We believe that theories of complex skill acquisition might help to develop performance assessments in open-ended game environments. We view the playing of a CBSG as a problem-solving process in which domain-specific prior knowledge (DSPK) has an essential role. DSPK comprises knowledge structures in long-term memory, also known as cognitive schemas.[13] Without these schemas, learners depend on domain-general problem-solving strategies which are inefficient and time-consuming and, most importantly, hamper the schema construction processes.[14] This means that playing a CBSG without sufficient DSPK might lead to suboptimal learning. The goal of this study is to empirically examine the effect of DSPK on game performance by comparing learners with two distinct levels of DSPK: learners with high DSPK (i.e., experts) and learners with low DSPK (i.e., novices). To this end, we develop specific assessment constructs and corresponding indicators, based on theories of complex skill acquisition.

In this introduction, we will first theoretically compare how experts and novices generate problem solutions, suggesting aspects of problem-solving performance that are directly affected by the level of DSPK. We will then discuss how to define indicators of these aspects by decomposing the skill structure hierarchically. Finally, we present the hypotheses of this study.

Figure 6. The model of problem solution generation by experts and novices



1.1. Problem Solution Generation by Experts and Novices

Figure 6 provides a process model that shows how experts and novices generate problem solutions differently, adopting concepts from the four-component instructional design (4C/ID) model.[15] The process involves two types of knowledge in long-term memory: *domain models* (i.e., schemas of how a domain is organized) at *declarative level*, and *cognitive strategies* (i.e., schemas of how to approach problems in the domain) at *procedural level*. Assume a continuum with novices with low DSPK at one extreme and experts with high DSPK at the other extreme.

For novices, since their domain models are not yet structured, *weak methods* (i.e., slow and inefficient general problem-solving strategies such as *general search* or *working backward*[16] are the only cognitive strategies they can use when solving a problem. This leads to *inefficient approaches* to the problem, and also to *procedures with incorrect cognitive rules* at the level of task performance (i.e., *solution generated*). For experts, on the other hand, well-structured domain models are interpreted and transformed into two types of stronger cognitive strategies: *knowledge-based methods* (i.e., heuristic strategies) and *strong methods* (i.e., algorithmic strategies).[14] Knowledge-based methods guide students to reason within the domain and systematically approach *non-routine* aspects of the problem (i.e., *systematic approach*). When a certain aspect of the given task is consistently repeated (i.e., *routine* aspects of tasks), cognitive if-then rules may be formed as strong methods. These rules provide algorithmic solutions to routine aspects of the task by matching conditional information in working memory (i.e., *if* part) with a coordinated reaction (i.e., *then* part), resulting in *procedures with correct cognitive rules* at the task performance level. As a function of extensive practice, the cognitive rules can be strengthened and eventually become fully automatized, leading to higher speed in performance.[17]

Additionally, the schemas embodied in long-term memory cause one more distinction in task performance between experts and novices: reduced cognitive load resulting from optimized use of working memory. Problem-solving with weak methods imposes a heavy demand on cognitive resources in working memory,[18] introducing high cognitive load or even cognitive overload.[19] However, with the availability of knowledge-based methods, cognitive schemas relevant for problem-solving are stored in long-term memory and retrieved into working memory as one element. Moreover, with fully automatized strong methods, cognitive schemas are activated directly without placing any demand on working memory resources, which further frees up working memory.[20]

Consequently, we derive four constructs that represent aspects of task performance that are affected by DSPK: (1) systematicity in task approach (i.e., representation of acquired strategies), (2) accuracy in applying cognitive if-then rules (i.e., representation of formed cognitive rules), (3) speed in performance (i.e., representation of the strength of those rules), and (4) reduced level of cognitive load (i.e., representation of optimized process).

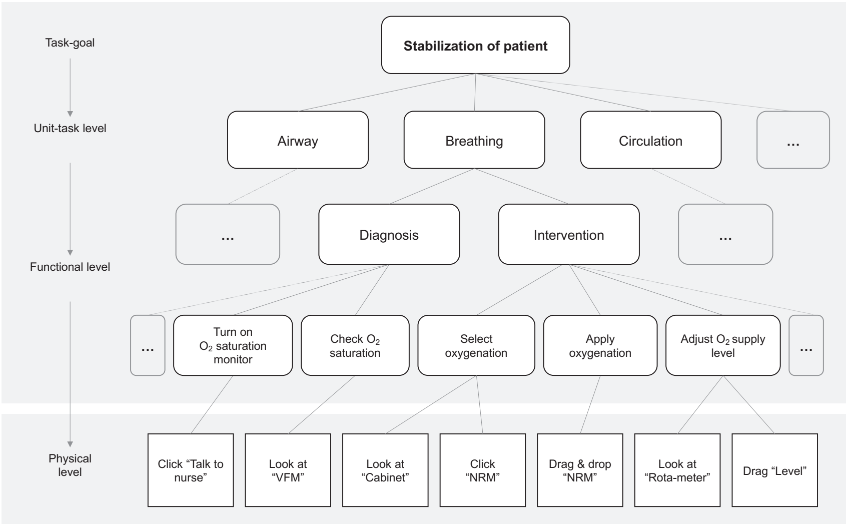
1.2. Skill Decomposition

While the four aspects of task performance are applicable to all task environments, *indicators* to assess these aspects will be specific to a particular task environment. Researchers have strongly recommended that, to assess a certain task performance, constituent skills and their relationships should be identified in a process of skill decomposition.[15,21] We deem that skill decomposition allows identification of the indicators of the four constructs mentioned above to be precise and theoretically sound.

The domain of this study is a resuscitation procedure, called the ABCDE method. The five letters ABCDE represent the five phases (i.e., Airway, Breathing, Circulation, Disabilities, Exposure) that a task performer goes through sequentially to stabilize an acutely ill patient. The sequence should be rigidly followed, based on the principle "treat first what kills first". AbcdeSIM[22], a CBSG for training the ABCDE method, is employed as the task environment. We decompose the task and develop a skill hierarchy by using Lee and Anderson's task analysis method[23,24] (Figure 7). In the hierarchy, the *task-goal* (i.e., stabilization of patient) is gradually divided into three levels: unit-task level, functional level, and physical level. To achieve the task-goal, *unit-tasks* (i.e., the five phases in the ABCDE method) are arranged accordingly. Each unit-task comprises multiple sub-tasks at the *functional level* (i.e., diagnosis and intervention). Every functional task is linked to individual activities at the *physical level* (e.g., look at "VFM", click "Talk to nurse").

What one can empirically measure is this physical level only, while other levels represent cognitive performance.

Figure 7. Skill hierarchy of AbcdeSIM.



Note. The task-goal is divided into three levels: unit-task level, functional level, and physical level.

This hierarchy guides us in the development of the indicators of the four constructs, by identifying different task levels. For the first construct (systematicity), a systematic approach in the ABCDE principle can be defined as a high level of adherence to the order of the five phases at the unit-task level. The challenge of measuring this construct is that the systematic approach is not directly observable at the physical level. To see this, note that the knowledge-based methods deal with non-routine aspects of a task, using the *same* knowledge *differently* based on rules-of-thumb.[14] An action that is associated with a certain ABCDE phase can also be taken during other phases strategically. Thus, an irregular ABCDE sequence observed at the physical level does not necessarily represent irregular performance at other levels. Consequently, the indicator of this construct should concern the hidden cognitive performance at the unit-task level, rather than analyzing the physical level only.

For the second construct (accuracy in applying cognitive rules), we recall that cognitive rules consist of *if* and *then* parts. In CBSGs in general, the *if* part emerges as information gathering via visual selection (i.e., looking at a particular part of the screen), while the *then* part corresponds to motor reaction to the task environment (e.g., mouse clicks or keyboard input). The accuracy in visual selection and motor reaction might reside at the functional level in the skill hierarchy. This construct can be directly detected by observing the physical level, because the strong methods deal with routine aspects of a task, referring to the *same* use of *same* knowledge.[14]

The third construct, the strength of cognitive rules, is situated in the connection between the sub-tasks at the functional level. If this connection is strong and stable, certain motor reactions at the end of a series of cognitive rules will be performed fast. The indicator of the strength should be the speed of this motor reaction, observed again at the physical level.

Lastly, the construct of reduced cognitive load originates from well-structured cognitive schemas. The entire structure of the skill hierarchy shows how the relevant cognitive schemas in long-term memory are developed, resulting in optimization of use of working memory. The degree of optimization can be indicated by the level of cognitive load.

1.3. Hypotheses

Four hypotheses will be tested in the current study:

H1 (systematicity in approach). Participants with high DSPK (i.e., experts) show higher systematicity in approach than participants with low DSPK (i.e., novices) during performance in the AbcedSIM, demonstrating higher level of adherence to the ABCDE phases.

H2 (accuracy in applying cognitive rules). Experts show higher accuracy in visual selection (H2a) and in related motor reactions (H2b) than novices.

H3 (speed in performance). Experts show higher speed in performance than novices.

H4 (reduced cognitive load). Experts experience lower cognitive load than novices.

Table 1 provides an overview of the constructs, hypotheses, and indicators. Details of indicators are described in the following method section.

Construct	Hypothesis	Indicators
Systematicity in approach (H1)	Experts adhere to the five ABCDE phases more rigidly than novices	· Hidden Markov model score that measures to what extent the order of the five phases was kept
Accuracy in visual selection (H2a)	Experts allocate more visual attention to critical diagnosis areas (CDA) than novices	· Proportion of dwell time in CDA · Proportion of fixation count in CDA · Proportion of fixation duration in CDA
Accuracy in motor reaction (H2b)	Experts complete functional tasks more accurately than novices	· Intervention completion score
Speed in performance (H3)	Experts complete functional tasks faster	· Time on intervention tasks
Cognitive load (H4)	Experts' richer and more automated schemas require less working memory capacity	· Cognitive load questionnaire score · Average fixation duration · Fixation frequency · Transition rate

2. Method

2.1. Participants and Design

Participants (*N* = 46) were recruited on a voluntary basis from a medical center in the Netherlands. For the expert group, residents in their second to fifth (final) year of residency training with an average 3.0 years of experience in emergency departments (*SD* = 2.1) were recruited (*n* = 24, mean age 31 years, *SD* = 4.2, 17 females). For the novice group, medical students in their second to sixth academic year who had been taught the basics of emergency medicine but had received no training were recruited (*n* = 22, mean age 23 years, *SD* = 2.0, 12 females). A causal-comparative design is adopted with the level of expertise as the single factor.

2.2. Material and Apparatus

2.2.1. Abcdesim Game Set-Up

AbcdeSIM is a medical simulation game to train the ABCDE method for resuscitation. The game starts with a storyline where users meet a virtual patient in an emergency room. The users are provided with tools for diagnosis (e.g., stethoscope, penlight) and intervention (e.g., infusion fluids, medication). Human physiology (e.g., respiration, circulation) of the patient is implemented in the game, giving feedback on user's interventions. A regular adult patient scenario, hemorrhagic shock due to gastrointestinal bleeding (GIB), was used. GIB is a scenario where learners should follow the basic ABCDE method, with most emphasis on the circulation phase. During the experiment, the game was run on a personal computer (Intel Core i7 2.67GHz CPU, 1.98 GB RAM) and presented on a Dell 22 "LCD screen with a resolution of 1650 x 1080 pixels. Participants used a headset for sound effects and interaction with the simulation was done via the mouse.

2.2.2. Eye-Tracking and Game-Log Recording

The game log data were saved in JSON file format (www.json.org). Participants' eye movements were measured by an SMI RED remote eye-tracker (SensoMotoric Instruments GmbH, Teltow, Germany) with a sampling rate of 250 Hz. The SMI Experiment Center 3.5 software (version 3.2.11, www.smivision.com) was used to implement calibration, validation, stimulus presentation, and screen recording. Eye movement data was gathered via SMI iView X software (version 2.7.13).

2.2.3. Cognitive Load Questionnaire

The NASA Task Load Index (NASA-TLX)[25] was used as a validated self-report questionnaire of cognitive load. It is a mental workload assessment tool for human-machine interaction domains such as aviation and aeronautics,[26] healthcare,[27] and socio-technical fields.[28,29] The NASA-TLX provides an overall workload score with six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Certain wordings of the questionnaire were adapted to fit the game environment.

2.3. Procedure

Individual sessions were carried out in an eye-tracking laboratory at Maastricht University. First, participants were asked to sign an informed consent form and fill out a questionnaire about demographics and experience in emergency medicine. Then, a pre-training was provided to ensure that the level of game-specific knowledge (i.e., how to operate the game) was comparable between the expert and novice groups. After pre-training, additional time for participants to play around with a test scenario was given, to allow them to familiarize themselves with the game. When participants expressed their readiness, the GIB scenario was presented. The eye-tracking system was calibrated, and validation followed directly. Participants had to stabilize the virtual patient in a maximum of 15 minutes, shown with a timer visible on screen during the entire session. After the scenario, participants filled out the NASA-TLX. The average time to complete a session was about 50 minutes.

2.4. Data Analysis

For testing H1, H2b, and H3, the data from game logs was used, while eye-tracking data was employed for testing H2a and H4. Parsing of the game logs was performed using Python. Eye-tracking data of three experts and two novices were excluded due to low tracking ratio below 85%. The average tracking ratio after the exclusion was 94.9%. Statistical analysis for each construct was performed in R version 3.5.1.[30]

2.4.1. Systematicity in Approach (H1)

We consider Hidden Markov Models (HMM) a suitable method to develop a score for measuring systematicity in approach, since they can be used to model *hidden state transitions* (i.e., phase arrangement at the unit-task level) based on a sequence of *emission states* (i.e., arrangement of motor reactions observed at the physical level).[31] The probability structure resulting from fitting the HMM to participant data contains information about the level of the adherence to the ABCDE sequence in hidden states. We used this probability structure to compute our score for systematicity in approach.

To do this, first, we classified the functional tasks of the GIB scenario into each of the ABCDE phases. Then, user-input data relevant to these functional tasks was extracted from the raw data in the game log file. The extracted data comprises the emission state sequences of ABCDE for each participant. A HMM is fitted to the sequences, resulting in a probability structure with two matrices: a state transition probability matrix and an emission probability matrix. From these matrices, we developed the HMM score by averaging the diagonal sum of the state transition matrix (TPM score) and the diagonal sum of emission probability matrix (Rho score). The R package "hmm.discnp" was used for these calculations.[32]

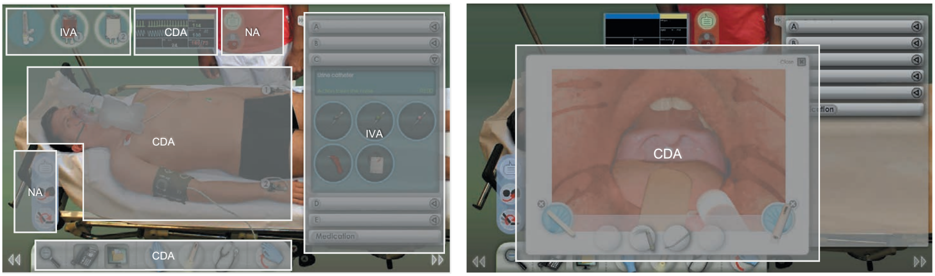
2.4.2. Accuracy in Visual Selection (H2a) and Motor Reaction (H2b)

Research in visual science reports that, compared to novices, experts allocate more attention to task-relevant than task-redundant areas.[33-35] In the AbcdeSIM, however, areas with information cannot simply be dichotomized as relevant versus redundant. Information and game functions are compactly organized within the limited area of the screen, and the level of relevance gradually differs. Thus, we categorized the areas on the screen into four groups in consultation with a medical professional: critical diagnosis area with critically relevant information for diagnosis (CDA), non-critical diagnosis area with information relevant for diagnosis to some extent but not critical (NDA), intervention area with functions for intervention (IVA), and neutral area with additional functions such as connecting different information (NA) (Figure 8). We hypothesized that experts allocate more attentional resources to CDA than novices, thus formulating H2a.

All area groups mentioned above comprised areas of interest (AOIs) forming the basis of the eye-tracking data analysis.[36] Since the appearance and layout of these areas dynamically change according to users' input and activated game function, we adapted the AOIs accordingly. The raw eye-tracking data was analyzed by SMI BeGaze 3.6 software. Fixations were identified when the gaze velocity was less than 40 visual degrees per second, with a minimum duration of 50 ms. Three eye-movement measures are employed: dwell time (gaze visiting time for an AOI from entry to exit), fixation count (number of fixations on an AOI), and fixation duration (time duration when the eye is relatively still). Each measure is expected to capture different aspects of attentional resources: time spent in the AOI, frequency of referencing the AOI, and robustness of processing the information in the AOI. To make the measures comparable across participants, relative values were calculated: the dwell time was divided by total play time, while the fixation count was divided by total number of fixations during the entire scenario. The fixation duration was divided by the fixation duration averaged over the scenario.

Visual selection and its associated motor reaction cannot be matched one-to-one, due to the dynamic characteristic of CBSGs. Thus, the accuracy in motor reaction was operationalized independently from the visual selection. We hypothesized experts complete more intervention tasks than novices, formulating H2b. The intervention completion score was developed as follows. In consultation with a medical professional, we selected five intervention tasks that are theoretically essential in the GIB scenario: *oxygen mask* *application, fluid administration, blood administration, blood order,* and *calling gastro-enterologists.* We then calculated the completion of these interventions using a binary measure and summed for each participant. This was done by extracting the corresponding data from the game log files.

Figure 8. Areas of interest (AOI) definition



Note. Critical diagnosis area (CDA), non-critical diagnosis area (NDA), intervention area (IVA), and neutral area (NA)

2.4.3. Speed In Performance (H3)

The relative time to complete the five intervention tasks from H2b was used as the speed measure. To make the time on each task comparable, z-scores were used. First, we checked whether the time-on-task per task was normally distributed. We then transformed each time-on-task into a z-score.

2.4.4. Cognitive Load (H4)

In addition to using the cognitive load questionnaire (i.e., NASA-TLX) as a subjective rating, we used eye-tracking measurements as an objective indicator of cognitive load. Several studies have shown that high cognitive load is related with long fixation durations[37,38] and high fixation frequency.[39-41] We also included transition rate (i.e., the movement from one AOI to another per second) that has been used in several studies of working memory capacity.[36] As cognitive load represents the level of optimization of working memory, we assume that a robust transition rate might be interpreted as an active cognitive process with an optimal use of working memory (i.e., a low level of cognitive load). Average fixation duration and fixation frequency were calculated over the scenario. Transition was counted using all individual AOIs from the four AOI groups aforementioned. Then the per-second transition rate was calculated.

3. Results

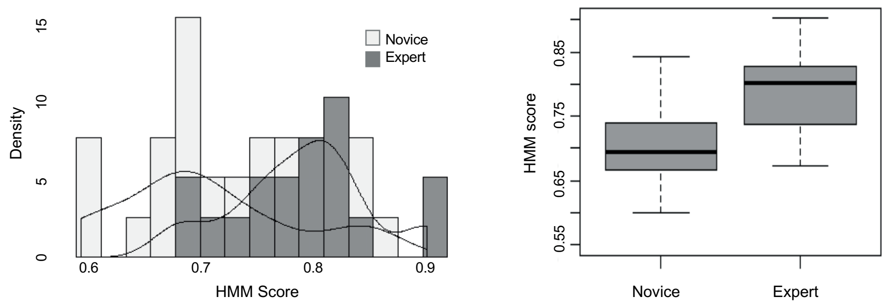
All measures for each construct were compared between experts and novices by t-tests for independent samples. When the data is not normally distributed, Mann-Whitney U test was used instead. MANOVA was used for comparing multivariate variables. Table 2 provides an overview of the outcomes of the variables related to all constructs except visual selectivity that is specified separately in Table 3.

Table 2. Correlation between cognitive load measures

Variable	1	2	3
1. NASA-TLX			
2. Average fixation duration	-.14		
	[-.43, .18]		
3. Fixation frequency	-.15	.39*	
	[-.44, .17]	[.09, .62]	
4. Transition rate	-.39*	.34*	.64**
	[-.62, -.09]	[.03, .58]	[.41, .79]

Note. Values in square bracket indicate the 95% confidence interval for each correlation.
* $p < .05$. ** $p < .01$.

Figure 9. Distribution of HMM score (left) and boxplot (right)



3.1. Systematicity in Approach

Figure 9 shows the distribution and the boxplot of HMM scores. The HMM score was significantly different between the two groups with a large effect size ($t(32) = -3.49$, $p = .001$; $d = -1.16$), indicating that experts adhered better to the ABCDE sequence than novices. There was no significant difference between groups in the length of the ABCDE sequences.

3.2. Accuracy in Visual Selection and Motor Reaction

Error! Reference source not found. demonstrates an overview of outcomes of visual selectivity measures for each AOI category. A MANOVA was conducted for all three relative measures of visual selectivity (i.e., dwell time, fixation count, and fixation duration) for CDA. The MANOVA revealed a significant difference between experts and novices ($F(3,37) = 4.67$, $p = .007$). Further, separate t-tests on the all three variables showed significant difference: experts showed higher proportion of dwell time to total play time with a large effect size ($t(38) = -2.62$, $p = .012$; $d = -0.82$), higher ratio of fixation count to total fixation counts with a medium effect size ($U = 123$, $p = .023$, $r = 0.35$), and longer fixation duration with a large effect size ($t(37) = -2.89$, $p = .006$; $d = -0.93$) for CDA. Different results were shown for the other three AOI categories. There was no significant difference for NDA and NA, except for the experts' higher fixation duration in NDA ($t(35) = -2.34$, $p = .025$; $d = -0.75$). As for IVA, experts showed significantly lower proportion in fixation count ($t(38) = 2.07$, $p = .045$; $d = 0.65$) and longer fixation durations ($t(35) = -2.43$, $p = .021$; $d = -0.79$). The intervention completion score was significantly higher, with a medium effect size, for experts than for novices ($U = 167.5$, $p = .027$; $r = 0.33$).

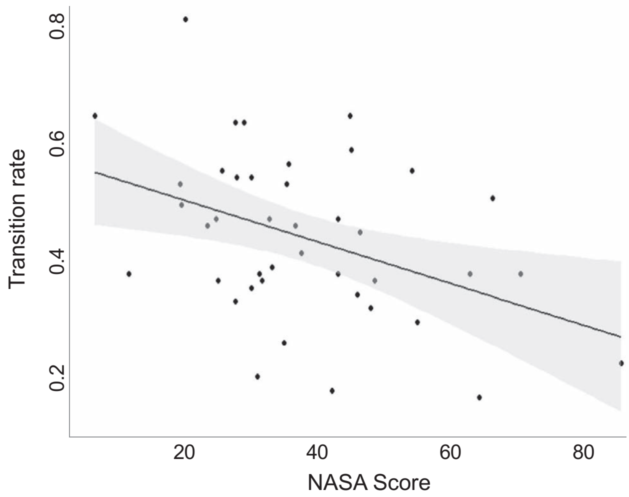
3.3. Speed

Experts showed faster unit-task reaction time with a large effect size ($t(40) = 3.77$, $p = .001$; $d = 1.14$). There was no significant difference in the total time on entire scenario performance.

3.4. Cognitive Load

A MANOVA was conducted for all four measures of cognitive load: NASA-TLX score, average fixation duration, fixation frequency, and transition rate. The MANOVA revealed a significant difference between experts and novices ($F(4,36) = 2.68$, $p = .047$) in cognitive load. However, separate t-tests showed diverged results. The NASA-TLX scores was lower for experts than novices, with a large effect size ($t(40) = 2.33$, $p = .025$; $d = 0.70$). There was no significant difference between experts and novices in average fixation duration ($t(36) = -0.41$, $p = .682$) and fixation count ratio ($t(36) = -0.30$, $p = .763$). Experts showed higher transition rate than novices, with a large effect size ($t(39) = -3.09$, $p = .004$; $d = -0.97$). Convergence between the NASA-TLX score and each eye-tracking measurement was examined via the Pearson correlation coefficient (Table 4). Transition rate displayed a negative correlation with the NASA-TLX scores ($r(39) = -0.39$, $p = .012$) (Figure 10), while other eye-tracking measures were correlated between themselves.

Figure 10. Scatter plot of NASA-TLX scores and transition rate



4. Discussion

This study aimed to empirically determine whether the level of domain-specific prior knowledge (DSPK) affects performance in a computer-based simulation game (CBSG). In the introduction, we argued that, to assess this performance, certain constructs should be developed by taking theories of complex skill acquisition as a starting point. We suggest four theoretical aspects of problem-solving performance to represent the level of DSPK, and defined indicators of these aspects based on a skill hierarchy, which resulted in four hypotheses. To confirm these hypotheses, game logs and eye-tracking data were collected and analyzed, using the methods corresponding to each aspect.

Hypothesis 1 stated that participants with high DSPK (i.e., experts) show higher systematicity in their approach during performance in a CBSG than participants with low DSPK (i.e., novices). The results of this study support this hypothesis. Systematicity for the AbcdeSIM task environment was defined as a high level of adherence of the ABCDE sequence at unit-task level, while flexibly adjusting task performance at physical level. According to the result from the HMM, the experts showed a higher level of adherence than novices to the ABCDE sequence at a hidden level. Additionally, the length of the ABCDE sequences at the physical level did not show significant difference between experts and novices. This implies that the important difference between experts and novices resides in the inner structure of the task performance, rather than the amount of physical action.

Hypothesis 2 concerns the accuracy in applying cognitive rules, stating that experts show the availability of more accurate cognitive rules. We decomposed the cognitive rules into two parts specific to CBSG environments: visual input of information from the environment (i.e., *if* part) and motor reactions to the environment (i.e., *then* part). Therefore, the hypothesis consisted of two sub-hypotheses: Experts show higher accuracy in both visual selection (H2a) and motor reaction (H2b) than novices.

H2a was confirmed, showing that experts have higher accuracy in visual selection of areas with critical information. This construct was operationalized as the ratio of allocation of visual attentional resources to critical diagnosis areas (CDA). All three eye-tracking metrics that were used (i.e. dwell time, fixation count, and fixation duration) indicated a higher allocation to CDA for experts. Additionally, we observed a variance between the three metrics for other areas, which can be explained by a psychological interpretation for each metrics. Dwell time indicates the time that a participant spent fixating on an AOI, where constituent metrics are not decomposed.[42] High number of fixation count indicates frequent reference to the stimulus,[43] while longer duration of fixations can mean a longer processing time.[44] For the areas with non-critical information for diagnosis (NDA), it was only fixation duration that showed a difference between experts and novices. This could mean experts and novices refer to minor diagnostic information by a similar amount, while experts process the same information deeper than novices. Interestingly, for the areas with intervention functions (IVA), novices showed higher fixation count, while experts showed longer fixation duration. These results suggest that novices search more frequently for what to execute (i.e., intervention) in the absence of an accurate diagnosis. This also can be interpreted as novices using weak methods such as *general search*[45] and *working backward*.[46] On the other hand, the experts seemed to process more deeply while not searching as much as novices in this area. Since the major components of IVA involve options of medical intervention tools, we suppose that experts choose intervention tools more carefully without redundant searching. Consequently, this leads experts to higher accuracy in their motor reactions, which supports H2b. The results showed that experts achieve higher scores in completion of essential interventions. Thus, Hypothesis 2 was largely confirmed: Compared to novices, experts allocate more visual attentional resources to critical information, followed by more appropriate motor reactions.

With regard to H3 which concerns speed in performance, the results confirmed that experts show higher speed than novices in task performance. Note that cognitive rules are formed via two sub-processes: proceduralization of abstract activities (i.e., organizing tasks at the functional level) and specialization of these activities to a particular physical environment (i.e., applying to the physical level).[47] Since the organization of functional tasks should be already stable for experts, the total strength from the two sub-processes would be enhanced for the experts. On the other hand, although experts demonstrated higher speed in performing specific unit-tasks that are most essential for the designated scenario, the entire performance duration showed no significant difference between experts and novices. Experts seem to complete the essential interventions faster, then repeatedly perform *reassessment* (i.e., repetition of the ABCDE procedure to ensure the process and monitor patients' changes), resulting in a similar length of overall performance time between experts and novices.

Lastly, Hypothesis 4 pertained to the lower level of cognitive load for experts compared to novices. This was supported in that experts reported lower cognitive load in a subjective rating scale (NASA-TLX), which was correlated with high transition rate. While subjective rating scales are a well validated measure of cognitive load, the interpretation of transition rates has been inconsistent in the literature. A robust transition rate can be related with optimal use of working memory[48-50] or better integration between different information sources,[51-53] which is on the same line with our interpretation. On the other hand, a high transition rate could also be connected to difficulties in integrating information sources,[54] inefficient visual problem-solving strategies,[55] or extraneous cognitive load caused by seductive details in multimedia learning.[56]

We presume that these different interpretations stem from differences in characteristics of visual stimuli which are highly task-dependent. When a task presents static stimuli with fixed information (e.g., static texts or figures), shifting eye-gazes between AOIs might indicate a stagnation within the same information. In this case, AOIs with information that has already been processed become irrelevant areas that does not require revisits.[55] On the other hand, when a task provides dynamically changing pieces of information, shifts between AOIs signifies rather a different kind, vigorous progress in gathering new information. Especially in medical simulations, monitoring patients' physiological changes and reacting upon them constantly (i.e., reassessment) is a major part in problem-solving, which can be facilitated by optimal use of working memory. Our result accords with the results of a previous study in ultrasound simulation,[57] which also used a medical simulation task. Furthermore, this dynamic aspect of medical simulation tasks seems to reduce the sensitivity of fixation duration and fixation count during the overall task performance in measuring cognitive load, due to the fluctuation of these measures. Further research is needed on using eye-tracking to measure cognitive load in different task environments.

The results of this study have several implications for indicator development in CBSGs. First of all, multiple aspects of performance should be considered as a whole when determining constructs to assess performance. Researchers in education have argued that a well-designed performance assessment should involve all aspects of performance containing more than one standard for each aspect, stating "one measure of performance is no measure of performance."[58,59] We suggest that the same principle should be applied to assessing task performance in CBSGs. This is the most important reason to use complex skill acquisition theories as a driving force, because it facilitates identification of different aspects of tasks as non-routine and routine, which assists further classification of the four aspects of performance. Secondly, since constructs are abstract and conceptualized broadly, they should be operationalized to concrete indicators that are fully designated to a specific task. This should be done through a task analysis in consultation with domain experts, also driven by educational theories. For instance, systematicity in approach is one of the broad concepts we explored in this study, which was problematic to operationalize. Thus, we robustly applied theories in complex skill acquisition and the relevant domain (i.e., the ABCDE method), to define the indicator of systematicity. Thirdly, this study opens the potential of combining eye-tracking data with game-log to quantify performance in CBSGs. Since most of CBSGs depend on visual stimuli, eye-tracking can be an important source to obtain a complete account of certain aspects of performance. In this study, we have found two indicators from eye-tracking (i.e., visual selection and cognitive load) that can possibly be used in CBSGs. Future research is needed to apply these findings to assessment and the development of support in CBSGs (e.g., student modelling to adaptively support individuals).

The insights from this study can help educators to assess students' performance in CBSGs and provide scaffolding to students with a low level of DSPK. For instance, they might focus on the different aspects in students' performance, then adjust the level of scaffolding to enhance each aspect. When the systematicity in approach is not high enough, instructors might stimulate the student to construct domain-specific knowledge and strategies (e.g., advise them to consult learning resources with relevant information). When a student concentrates on reactions only without sufficient information gathering, they can guide the student to pay more attention to information gathering as a sound foundation for taking actions in the game. Additionally, when the student's cognitive load is high, extra support could be given to manage the load. This can be done either by reducing the cognitive load itself (e.g., providing pauses during the game or presenting less complex scenarios), or facilitate self-regulation of students to manage their own cognitive load.[20]

Several limitations of this study need to be mentioned. Firstly, our findings might not be generalizable to other CBSG environments since the indicators were specialized for a specific task. Future research should follow to examine how our methods can be applied in other CBSG environments. Secondly, the participants in the expert group were composed of residents, rather than medical doctors. In this study, we selected medical students as novices and residents as experts, in order to form comparable groups. Although it led to a better controlled experimental setup, including a wider range of expertise levels could have yielded more informative results. Thirdly, although one could well argue that eye-hand coordination in performing cognitive rules is another aspect of performance, it was not explored in this study. We rather analyzed visual selection and motor reaction separately based on our assumption that those two cannot be matched one-to-one in a dynamic environment of a CBSG. However, investigating eye-hand coordination as an aspect of performance via non-linear analysis should be an intriguing topic for future study.

In conclusion, this study has demonstrated the development of performance assessment that can be used in a highly dynamic game environment. This was accomplished by starting from theories of complex skill acquisition, identifying constructs for assessment and valid indicators. We believe empirical investigation for reliable indicators in CBSGs can be seen as a problem-solving process itself. As in problem-solving by experts, the research should be driven by a certain knowledge structure (e.g., educational theories) to avoid an inefficient process and suboptimal solutions. Educational theories and empirical experiment are in a reciprocal relationship, where one cannot stand alone without the other.

Appendix. The HMM score computation

We start computing the HMM score by first extracting the ABCDE phases of the subsequent observed actions from the log file. Next, a HMM is fitted to this sequence using an EM algorithm, as provided by the R package hmm.discnp.[32,60] The HMM is set to have 5 inner states (actual phases) and 5 emission values (observed actions). Since fitting an HMM using a single observed sequence is strongly dependent on the starting condition, we initialize the HMM with a transition matrix with most probability mass concentrated on the diagonal and upper co-diagonal, and an emission probability matrix with most probability mass concentrated on the diagonal.

The resulting probability structure after fitting to the observed sequence contains information on the adherence to the ABCDE order. In the transition probability matrix, the total probability on “forbidden” transitions (e.g., jump from A to E) show how much is deviated from the order. The probability on “forbidden” emissions (e.g., an action for B in phase D) in the emission probability matrix shows how often actions are taken in a wrong phase. From this obtained probability structure, we compute a score: the total probability on legal transitions plus the total probability on legal emissions, divided by 2. This score ranges between 0 and 1.

Consequently, the HMM score increases when a performer keeps to the ABCDE phases in order, while the score decreases when the performance deviates from the order. For instance, in case of an ideal performer, the hidden sequence follows the ABCDE phases in a complete order (e.g., A-A-A-A-A-B-B-B-C-C-C-C-D-D-D-D-D-E-E-E-E-E). The HMM score for this example is 1.0. In the case of a less ideal performer, the sequence may deviate from the complete order (e.g., A-A-A-B-E-C-C-E-D-B-C-A-C-B-D-C-E-C-D-E-D-A-E-E), signifying that this performer jumped around the phases using less ideal rules. The HMM score for this example is 0.792.

References

1. Aldrich C. Clark Aldrich's six criteria of an educational simulation. 2004.
2. Kiili K. Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education.* 2005;8(1):13-24. https://doi.org/10.1016/j.iheduc.2004.12.001
3. De Freitas S. Learning in immersive worlds: A review of game-based learning. 2006.
4. Hernández-Lara AB, Perera-Lluna A, Serradell-López E. Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Comput Human Behav.* 2018.
5. Dankbaar ME, Alsma J, Jansen EE, van Merrienboer JJ, van Saase JL, Schuit SC. An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. *Advances in Health Sciences Education.* 2016;21(3):505-521.
6. Kang J, Liu M, Qu W. Using gameplay data to examine learning behavior patterns in a serious game. *Comput Human Behav.* 2017;72:757-770. https://doi.org/10.1016/j.chb.2016.09.062
7. Squire K. Open-ended video games: A model for developing learning for the interactive age. *The ecology of games: Connecting youth, games, and learning.* 2008:167-198.
8. Wallner G, Kriglstein S. Visualization-based analysis of gameplay data–a review of literature. *Entertainment Computing.* 2013;4(3):143-155.
9. Loh CS, Sheng Y, Ifenthaler D. Serious games analytics: Theoretical framework. In: *Serious games analytics.* Springer; 2015:3-29.
10. Zhou M, Xu Y, Nesbit JC, Winne PH. Sequential pattern analysis of learning logs: Methodology and applications. *Handbook of educational data mining.* 2010;107.
11. Clark DB, Martinez-Garza MM, Biswas G, Luecht RM, Sengupta P. Driving assessment of students' explanations in game dialog using computer-adaptive testing and hidden Markov Modeling. In: *Assessment in game-based learning.* Springer; 2012:173-199.
12. Hou H-T. Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. *Comput Human Behav.* 2015;48:424-435.
13. Bartlett FC. Remembering: An experimental and social study. *Cambridge: Cambridge University.* 1932.
14. Van Merriënboer JJ. Perspectives on problem solving and instruction. *Computers & Education.* 2013;64:153-160. 10.1016/j.compedu.2012.11.025
15. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
16. Newell A, Simon HA. *Human problem solving.* Vol 104: Prentice-Hall Englewood Cliffs, NJ; 1972.
17. Palmeri TJ. Theories of automaticity and the power law of practice. 1999.
18. Sweller J, Clark R, Kirschner P. Teaching general problem-solving skills is not a substitute for, or a viable addition to, teaching mathematics. *Notices of the American Mathematical Society.* 2010; 57(10):1303-1304.

19. Sweller J. Cognitive load during problem solving: Effects on learning. *Cognitive science.* 1988; 12(2):257-285.

20. Sweller J, van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev.* 2019:1-32.

21. Gagne RM. Learning hierarchies. *Educ Psychol.* 1968;6(1):1-9.

22. *AbcdeSIM* [computer program]. Rotterdam: VirtualMedSchool; 2012.

23. Lee FJ, Anderson JR. Does learning a complex task have to be complex?: A study in learning decomposition. *Cognitive psychology.* 2001; 42(3): 267-316. https://ac.els-cdn.com/S0010028500907473/1-s2.0-S0010028500907473-main.pdf?_tid=1b-63cd9d-fefb-49ad-a6d3-2362d5add161&acdnat=1531744697_69c192ef3cf56ec1151a7c-2871d799a6.

24. Card S, Moran T, Newell A. The Psychology of Human-Computer Interaction. Hillsdale, New Jersey: Lawerence Erlbaum Associates. In: Inc; 1983.

25. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: *Advances in psychology.* Vol 52. Elsevier; 1988:139-183.

26. Shamo MK, Dror R, Degani A. A multi-dimensional evaluation methodology for new cockpit systems. Paper presented at: Proceedings of the Tenth International Symposium on Aviation Psychology1999.

27. Weinger MB, Vredenburgh AG, Schumann CM, et al. Quantitative description of the workload associated with airway management procedures. *J Clin Anesth.* 2000;12(4):273-282.

28. Warm JS, Matthews G, Finomore Jr VS. Vigilance, workload, and stress. In: *Performance under stress.* CRC Press; 2017:131-158.

29. Grigg SJ, Garrett SK, Benson LC. Using the NASA-TLX to Assess First Year Engineering Problem Difficulty. Paper presented at: IIE Annual Conference. Proceedings2012.

30. *R: A language and environment for statistical computing* [computer program]. R Foundation for Statistical Computing; 2010.

31. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics.* 1970;41(1): 164-171.

32. Turner R, Liu L. hmm. discnp: Hidden Markov models with discrete non-parametric observation distributions. In: R Package Version 0.2-4, URL https://CRAN. R-project. org/package= hmm. discnp; 2016.

33. Reingold EM, Sheridan H. Eye movements and visual expertise in chess and medicine. *Oxford handbook on eye movements.* 2011;528.

34. Gibson JJ. *The ecological approach to visual perception: classic edition.* Psychology Press; 2014.

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

36. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J. *Eye tracking: A comprehensive guide to methods and measures.* OUP Oxford; 2011.

3

37. Korbach A, Brünken R, Park B. Learner characteristics and information processing in multimedia learning: A moderated mediation of the seductive details effect. *Learning and Individual Differences.* 2016;51:59-68.

38. Park B, Knörzer L, Plass JL, Brünken R. Emotional design and positive emotions in multimedia learning: An eyetracking study on the use of anthropomorphisms. *Computers & Education.* 2015; 86:30-42.

39. Van Orden KF, Limbert W, Makeig S, Jung T-P. Eye activity correlates of workload during a visuospatial memory task. *Hum Factors.* 2001;43(1):111-121.

40. Van Orden KF, Nugent W, La Fleur B, Moncho S. *Assessment of variable coded symbology using visual search performance and eye fixation measures.* NAVAL HEALTH RESEARCH CENTER SAN DIEGO CA;1998.

41. Zelinsky GJ, Rao RPN, Hayhoe MM, Ballard DH. Eye movements reveal the spatiotemporal dynamics of visual search. *Psychol Sci.* 1997;8(6):448-453.

42. Orquin JL, Holmqvist K. Threats to the validity of eye-movement research in psychology. *Behav Res Methods.* 2018;50(4):1645-1656.

43. Orquin JL, Loose SM. Attention and choice: A review on eye movements in decision making. *Acta Psychol (Amst).* 2013;144(1):190-206.

44. Rayner K. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin.* 1998;124(3):372.

45. Gick ML. Problem-solving strategies. *Educ Psychol.* 1986;21(1-2):99-120.

46. Larkin J, McDermott J, Simon DP, Simon HA. Expert and novice performance in solving physics problems. *Science.* 1980;208(4450):1335-1342.

47. Anderson JR. *How can the human mind occur in the physical universe?* : Oxford University Press; 2009.

48. Epelboim J, Suppes P. A model of eye movements and visual working memory during problem solving in geometry. *Vision Res.* 2001;41(12):1561-1574.

49. Miall RC, Tchalenko J. A painter's eye movements: A study of eye and hand movement during portrait drawing. *Leonardo.* 2001;34(1):35-40.

50. Berséus P. Eye movement in prima vista singing and vocal text reading. 2002.

51. Bartels M, Marshall SP. Eye tracking insights into cognitive modeling. Paper presented at: Proceedings of the 2006 symposium on Eye tracking research & applications2006.

52. Schmidt-Weigand F, Kohnert A, Glowalla U. A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and instruction.* 2010;20(2):100-110.

53. Johnson CI, Mayer RE. An eye movement analysis of the spatial contiguity effect in multimedia learning. *J Exp Psychol Appl.* 2012;18(2):178.

54. Holsanova J, Holmberg N, Holmqvist K. Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition.* 2009;23(9):1215-1226.

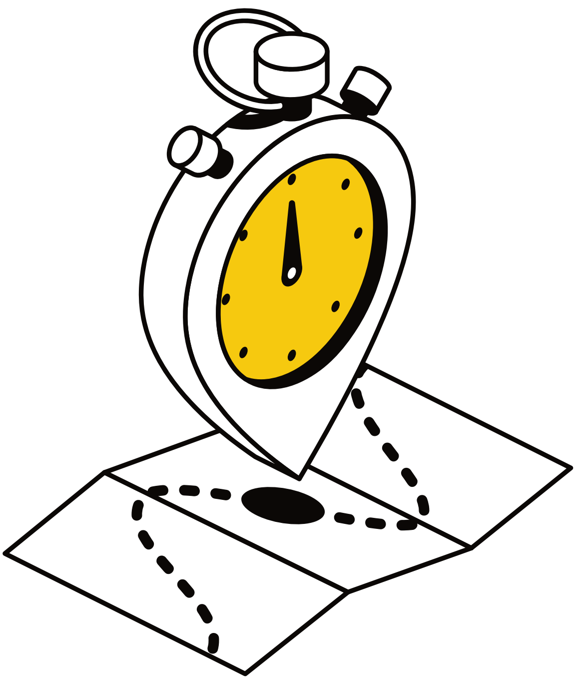
55. van Meeuwen LW, Jarodzka H, Brand-Gruwel S, Kirschner PA, de Bock JJ, van Merriënboer JJ. Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction.* 2014;32:10-21.

56. Korbach A, Brünken R, Park B. Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instructional science.* 2017;45(4):515-536.
57. Aldekhyl S, Cavalcanti RB, Naismith LM. Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on medical education.* 2018:1-10.
58. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA.* 2009;302(12):1316-1326.
59. Baartman LK, Bastiaens TJ, Kirschner PA, Van der Vleuten CP. The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in educational evaluation.* 2006;32(2):153-170.
60. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989 Feb;77(2):257-86.

Chapter 4

Different Effects of Pausing on Cognitive Load in a Medical Simulation Game

This chapter is published as:
Lee JY, Donkers J, Jarodzka H, Sellenraad G, Van Merriënboer JJG. Different effects of pausing on cognitive load in a medical simulation game. Comput Human Behav. 2020;110:106385. 10.1016/j.chb.2020.106385



Abstract

In medical training, allowing learners to take pauses during tasks is known to enhance performance. Cognitive load theory assumes that insertion of pauses affects cognitive load positively and thus enhances performance. However, empirical studies on how allowing and taking pauses affects cognitive load and performance in dynamic task environments are scarce. We investigate the pause effect, using a computerized simulation game in emergency medicine. Medical students (N = 70) were randomly assigned to two conditions: simulation with the option to take pauses available (n = 40), and simulation without the pause option (n = 30). All participants played the same two scenarios, during which game logs and eye-tracking data were recorded. Overall, both cognitive load and performance were higher in the pause-available condition than in the pause-unavailable condition. The act of pausing, however, lowered cognitive load temporarily, especially during intense moments. Two different manifestations of the pause effect are identified: (1) availability of pauses increases overall cognitive load by stimulating additional cognitive and meta-cognitive processes, (2) execution of pausing temporarily decreases heightened cognitive load through relaxation. Conclusively, our results present that it is important to guide students to utilize different effects of pausing to enhance their performance and learning.

Keywords: Cognitive load, medical education, pause effect, medical simulation, serious game, eye-tracking

1. Introduction

'Time-outs', where learners take pauses during their task performance to provide a moment for reorientation, are well-known as an effective technique to enhance performance in medical training.[1-3] The pauses can be proactively planned as a preoperative briefing where performers reflect on the current diagnosis and ensure that certain parameters have been met,[4] or be triggered by an unusual issue (e.g., recognition of unusual anatomy, a patient's unexpected reaction).[2] Furthermore, it has been reported that employing time-outs in surgery substantially improves patient safety.[1] Clinical educators have been promoting the use of time-outs as one of the essential instructions for trainees to enhance their performance and learning.[3]

Although the pause technique has received some attention in medical education, little insight into how pauses enhance performance has been established within a theoretical framework, nor are the insights supported by sufficient evidence from empirical research. The present study extends the research span of the pause effect in two ways: (1) developing hypotheses about the pause effect by employing recent descriptions of cognitive load theory, such as its application in medicine[5] and the connection to self-regulation,[6] and (2) verifying the pause effect empirically by using a computerized medical simulation game and objective physiological measures such as pupillometry.[7]

1.1. Cognitive Load Theory in Medical Education

Cognitive load theory (CLT) is a well-developed framework to explain the relation between instructional design, cognitive load, and learning processes. The main idea of CLT is that various elements in instructional design evoke cognitive processes that impose loads on working memory, and the learning processes can be optimized only when the total sum of the loads does not exceed working memory capacity.[8] There are three types of cognitive load: *intrinsic load*, which represents the inherent demands of a task given the level of expertise of the task performer; *extraneous load*, which refers to a detrimental load from distracting aspects of the instructional environment, and *germane load*, which pertains to the mental effort directly invested in learning.[9,10] CLT presumes that an optimal instructional design should keep intrinsic load to a manageable size for individual students' capacity, minimize extraneous load, and maximize germane load, while still preventing overload.[11]

In medical education, CLT offers promising applications since medicine is a highly complex domain where multiple sets of knowledge and skills should be processed interactively in working memory.[12] Since these multiple elements should be grounded at a physical time and place in medical training, CLT researchers in medical education stress not only learning but also effective performance of authentic tasks.[11,13] Clinical researchers found that cognitive overload heavily degrades task performance in both simulation and clinical practice, also described as "helmet fire"[5,7] Fraser et al.[14] explored the relationship between cognitive load and diagnostic performance in high-fidelity simulation training. They argued that heightened emotions impose extraneous load while dynamic interaction with simulators escalates intrinsic load, finding that performance declines as the total cognitive load becomes too high.

1.2. Assumptions for the Pause Effect

To counteract "helmet fire", CLT researchers suggested specific strategies for medical training, such as segmenting, pretraining, simple-to-complex sequencing, and emotional management.[11,15,16] Inspired by the concept of time-outs in medicine, we focus on *pausing*, which we define as a *possible and voluntary action* that interrupts a given task performance to provide *extra time for necessary cognitive processes*. We expect that this pausing should enhance medical students' performance by positively affecting cognitive load in highly complex learning environments such as simulation training.

In the past, studies concerning the pause effect in CLT have yielded inconsistent results.[17-22] We argue that these studies have a limited explanatory power for the pause effect as defined above, as they all used relatively passive stimuli (e.g., learning from text and pictures, instructional animations) where learners' interaction with the task environment is extremely limited. Moreover, the pause effect in those studies is often blended with the *segmenting effect,*[23] because the pauses occur where the learning content is already logically segmented by the researchers.[23-26] In medical simulation training, however, the task environment is highly dynamic and interactive, where neither educators nor learners can make a clear structure of a given content a priori. We, therefore, develop two assumptions about the pause effect based on CLT, with regard to the two major aspects of pausing as defined above: (1) pausing as a possible and voluntary action, and (2) pausing as insertion of extra time for necessary cognitive processes.

The first aspect, pausing as a possible and voluntary action, signifies that pausing involves not only the act of pausing but also the arrangement of the act, given that the pauses are available. To take pauses in medical simulation, learners need to monitor their progress while taking the remaining time into account and decide for themselves when to pause and for what purpose, often in the middle of intense situations. This can be described by the concept of self-regulation which includes two subprocesses: self-observation that refers to learners' self-recording of their personal events, and self-control that refers to learners' deployment of specific strategies. According to Zimmerman,[27] there are three phases in self-regulation: forethought phase, performance phase, and self-reflection phase. The self-observation and self-control are included in the performance phase, which implies that these two subprocesses continue throughout the performance, real-time.

Self-regulation in general is beneficial to performance enhancement since learners can use their self-awareness to optimize their cognitive processes.[27] However, it requires learners to invest metacognitive resources, causing additional cognitive load.[28] Note that this particular cognitive load might be experienced even when learners never actually take pauses, as the self-observation and self-control apply to the performance phase in self-regulation. This would result in an increase of cognitive load at the overall level during performance, while performance itself is positively affected. Therefore, we arrive at the first assumption: availability of taking pauses would enhance performance through the increase of beneficial cognitive load caused by self-observation and self-control during the performance, as long as the learner's working memory is not overloaded.

The second aspect, pausing as insertion of extra time for necessary cognitive processes, allows for more complex assumptions since the necessary cognitive processes can be diverse. Considering that the time-outs in medicine take place either by proactive plan or as a response to emerging issues,[2] we identify two types of activities that can occur during pauses in medical training: *reflection* and *relaxation*. The pauses for reflection can be carried out to approach a structured review of one's diagnosis by recalling knowledge from long-term memory and integrating the knowledge with present conditions, also described as "looking at the textbook on the ceiling"[2] Since these cognitive processes are relevant to elaboration in learning, reflective pauses would cause an increase in germane load.

This reflective pause can also be described by the concept of self-evaluation, which is a subprocess of the self-reflection phase in self-regulation.[27] This self-evaluation is distinct from the self-observation in the performance phase. While self-observation refers to *reflection-in-action* that is prerequisite for self-control, self-evaluation pertains to *reflection-on-action* where learners assess their own performance with regard to the goals of the given task and available resources to accomplish these goals.[27,29] Since this self-evaluation, as a subprocess of self-regulation, boosts metacognitive processes, the pauses for reflection would increase cognitive load.

The pauses for relaxation can take place by unusual issues that arise during simulation. When a stressful and unexpected event occurs, learners can take pauses to relax and to take a deep breath. Since the pause prevents the influx of new information, learners can manage stress and its negative impact on working memory without being disturbed further. This type of relaxing pauses would reduce extraneous load arising from heightened emotions.

One can question how the different cognitive activities during pauses would influence cognitive load and performance, given that the availability of pauses was assumed to affect cognitive load and performance at the overall level. On the one hand, the activities might be influential only during the pauses, assuming that their effects on the overall procedure are weak. On the other hand, the activities during pauses could also be influential at the overall level. The reflective pauses stimulate the use of knowledge from long-term memory, and the self-evaluation of the current status with regard to the goals of the given task allows for a "fresh look"[3] on the current diagnosis. This may facilitate a structured approach to the task after the reflective pauses, resulting in a better use of working memory during performance (i.e., decrease in cognitive load) and performance enhancement at the overall level. Also, the relaxing pauses can maintain working memory in an optimized state during performance, by managing cognitive load whenever it soars high. This can prevent performance from deteriorating further, enhancing performance at the overall level.

To sum up, we propose the second assumption: execution of pausing can enhance performance by affecting cognitive load either during pauses only or at the overall level. The cognitive load during pauses can vary, depending on the type of cognitive processes that take place during the pause: it can increase by reflection-on-action, or decrease by relaxation. If the cognitive processes during the pauses affect the overall procedure, cognitive load would decrease and performance would enhance, since working memory is refreshed and optimized through the pauses.

1.3. Present Study

Based on the aforementioned considerations, this study aims to empirically investigate the pause effect on cognitive load and performance. We use a computerized medical simulation game in the domain of emergency medicine. The tasks in emergency medicine present highly complex and intense scenarios, where the pause effect is expected to be salient. Cognitive load is assessed by two types of measures: rating scales as a subjective psychometric measure, and pupillometry as an objective physiological measure. These two measures are known to be reliable cognitive load indices in medical simulation settings.[7,30,31] The use of different types of measures allows for a rich interpretation of different types of cognitive load.[32] Performance is measured by the in-game scoring system which is a checklist of correct interventions. Whether a learner properly managed an intense and stressful situation is also considered as a separate score.

We formulate five hypotheses divided into two categories, based on the two assumptions described above. Following the first assumption about the pause as a possible and voluntary action, the mental effort for self-observation and self-control can affect cognitive load at the overall level, regardless of whether or not a pause was actually taken during the game. If the total cognitive load (i.e., sum of the three types of cognitive load) is not overloading working memory, performance would enhance thanks to the activation of

beneficial cognitive processes. Thus, we hypothesize a directional effect of *allowing* pauses on cognitive load and performance:
H1a. Allowing learners to take pauses increases cognitive load at the overall level.
H1b. Allowing learners to take pauses enhances performance.
Regarding the second assumption about the time-insertion aspect, the act of taking pauses can affect cognitive load during pauses differently depending on the diverse cognitive processes that can take place during the pauses (i.e., reflection or relaxation). These cognitive processes can be influential either for the overall procedure or for the pause duration only. If the overall procedure is affected, the processes will decrease cognitive load and enhance performance. If only the pause itself is affected, the cognitive load during the pauses would differ from the cognitive load during game-play. Accordingly, we construct two hypotheses with directional effects of *taking* pauses on cognitive load and performance at the overall level, and one hypothesis with non-directional effects of taking pauses on cognitive load during pauses:
H2a. Given that pausing is allowed, the act of taking pauses decreases cognitive load at the overall level.
H2b. Given that pausing is allowed, the act of taking pauses enhances performance.
H2c. Cognitive load during pauses differs from the cognitive load during game-play.
The structure in these hypotheses will be reflected in the presentation and discussion of the results.

2. Method

2.1. Participants and Design

A total of 70 medical students from Maastricht University, the Netherlands (51 females; mean age 22.7; *SD* = 2.78) participated on a voluntary basis. Since the basics of emergency medicine are taught in the 1st year, the participants' academic years ranged from 2nd to 6th (the last academic year for medicine). Experience in emergency medicine varied depending on their individual curriculum, except the basic course in the 1st year. A between-subjects design is adopted with availability of pausing (i.e., existence of a pause button in the game) as the single factor. The participants were randomly assigned to two groups: pause-available condition (PA) where students are allowed to take pauses by pressing the pause button, and pause-unavailable condition (PU) where no pause button exists. Ten more students were assigned to the PA condition (n = 40) than the PU condition (n = 30), for a more comprehensive analysis within the PA condition where a distinction can be made between students who actually took pauses and students who did not.

2.2. Materials and Apparatus

2.2.1. Abcdesim Game

We used AbcdeSIM,[33] a computer-based simulation game to train emergency medicine. This game is designed to replicate real-life scenarios in a medical emergency department. In each scenario, an acutely ill patient is delivered to the emergency room where a nurse is assisting the procedure. Learners should stabilize the patient within 15 minutes, by diagnosing (e.g., laboratory testing, ordering ECG, using stethoscope), intervening (e.g., oxygen application, fluid infusion, medication), and monitoring the patient's vital signs (e.g., vital function monitor, reassessment). Since using the nurse's assistance is crucial in this game, learners should communicate with the nurse actively. A high-fidelity model of human physiology reacts to learners' interaction in real time, simulating dynamic situations in the emergency room.

Two scenarios were selected in consultation with two medical professionals, following a series of pilot studies: subarachnoid bleeding (SAB) and gastrointestinal bleeding (GIB). The complexity level of these scenarios is high enough for high cognitive load to be prominent. The SAB scenario presents a 46-year-old woman with a partially obstructed airway and epilepsy, while the GIB scenario presents a 32-year-old man with hypovolemic shock. In the SAB scenario, the epileptic seizures occur randomly and require treatment within an extremely limited time window, which creates a highly intense and stressful situation for the learner. If the learner has been playing the game systematically (e.g., prioritizing more urgent issues) before the epilepsy occurs, certain advantages in the treatment of the epilepsy are granted (e.g., more time for providing the treatment). In the GIB scenario, the patient's blood pressure drops fast continuously, so learners should take measures quickly and accurately.

The game provides the option of pausing which can be preset by the researcher. In the PA condition, a pause button is displayed besides a timer on the upper side of the screen. When the pause button is pressed, the screen is stopped as it currently is, and the timer also stops. The paused image is slightly greyed out with a message box saying "AbcdeSIM is paused" popping up, while it is still possible to see important information on the screen (Figure 11). There is no penalty in game score for using pauses. The simulation was operated on a personal computer (Intel Core i7 2.67 GHz CPU, 1.98 GB RAM), presenting the scenarios on 1650 × 1080 LCD monitor. Participants interacted with the game using a wireless mouse, and received sound effects from the scenarios (e.g., patients' abnormal snoring, lung sound, alarm from vital function monitor) using a headset.

Figure 11. Screenshots of AbcdeSIM in the PA (pause-available) condition.

Note. On the left side, a game-screen during performance is shown. A pause button is displayed besides a timer on the upper side of the screen. On the right side, a game-screen during a pause is shown. During the pause, the screen is slightly greyed out, while it is still possible to see important information.

2.2.2. Cognitive Load Rating Scales

A self-report rating for perceived mental effort[34] was used to measure the subjective overall cognitive load. Participants gave the rating right after finishing each scenario, using a sliding indicator with scores on a scale of 0 to 100.

2.2.3. Eye-Tracking and Game-Log Recording

A laptop computer (Intel Core i7 2.67 GHz CPU, 1.98 GB RAM) was designated for eye-tracking data recording, composing a dual-computer setup with the personal computer for game presentation. Eye-tracking data was recorded by an SMI RED remote eye-tracker at a sampling rate of 250 Hz, using SMI iView X software (version 2.7.13, www.smivision.com). The SMI Experiment Center 3.5 software (version 3.2.11) was used to arrange calibration, stimulus presentation, and screen recording. Since head-movements can distort the pixel size of the pupil in the remote eye-tracking system, participants were positioned in a forehead-and-chin rest and the geometry between the apparatus and the eye was preset, for stable calculations of pupillometry.[35] The experiment room was windowless, dimly lit, and quiet.

The eye-tracking raw data includes samples of pupil diameter, gaze point, and timestamps, while the game log data includes in-game events (e.g., taking pauses, onset of epilepsy) and timestamps. The game system was programmed to send synchronization signals to the iView X software, by which eye-tracking data and game logs could be merged and aligned.

2.3. Procedure

The experiment was carried out in individual sessions in an eye-tracking laboratory. To begin with, participants signed an informed consent form and filled out a demographic questionnaire. Then, they played through a tutorial of the game to familiarize themselves with the game functions. For the pause available condition (PA) group, the participants were instructed to take pauses whenever the need arises. After participants positioned themselves in the eye-tracking setup, calibration with a 9-point procedure followed. To establish a baseline for the individual, a randomly scrambled version of a game image was presented for 5 seconds.[35] In the PA condition, a baseline for the paused screen was also recorded since the paused screen had a different luminosity, due to the greyed-out effect. Starting a scenario, a maximum of 15 minutes was given to stabilize a patient, with a timer visible on the screen during the entire scenario. In the PA condition, a maximum of 10 pauses was available, and the number of remaining pauses was marked inside the pause button. After completing a scenario, participants rated the overall cognitive load. In the PA condition, a simple questionnaire about the participants' activities during the pauses was added. The procedure from calibration to the questionnaire composes one trial. The trial was repeated for the two scenarios (i.e., SAB and GIB), while the order of the scenarios was randomized. The entire session including the two trials lasted about 50 minutes.

2.4. Data Analysis

We compare PA against PU condition to test H1 (i.e., the effects of allowing pauses). With regard to H2a and H2b (i.e., the effects of taking pauses on cognitive load and performance at the overall level), we divided the participants in the PA group into two subgroups: PAn where participants take one or more pauses, and PA0 where no pauses are taken, then PAn is compared against PA0. The central values of all dependent variables were calculated per trial.

With regard to H2c (i.e., the effect of taking pauses on cognitive load during the pauses), we only use pupillometry data in PA as the dependent variable, as this measure continually records pupil dilation during performance. We sliced the timeline of scenarios in PA into two types of time periods: *game-play periods* where the scenario is running, and *pause periods* where the scenario is paused, then the pause periods were compared against the game-play periods. In addition, we take the different cognitive activities during the pauses (i.e., reflection-on-action or relaxation) into account. We assumed different levels of intensity of a given situation would produce different activities. Two different situations within a scenario were identified: *intense mode* where the patient has triggered an epileptic seizure, and *plain mode* where the patient is not having a seizure. All in all, the pause and game-play periods were compared separately in each of the two modes (i.e., intense and plain).

For statistical analysis of the pause effects over entire scenarios, we constructed linear mixed effect models, using the lme4 package[36] in R version 3.5.1.[37] As fixed effects, we entered group condition and scenario (without interaction term) into the model, while the participant was treated as a random factor. For the intensity-care score which takes on a binary value, a logistic model as a form of generalized linear mixed effect model was used. Since the intensity-care score is only available in SAB scenario, trials in GIB scenario were removed from the dataset, excluding scenario from the fixed factors. There were no evident deviations from homoscedasticity or normality from visual probe of residual plots. We calculated *p* values via likelihood ratio tests that compare the full model with the effects in question against the model without the effects in question.

2.4.1. Pupillometry Analysis

The merged data set (i.e., eye-tracking data and game logs) was imported into R for data processing. We controlled for three major confounding factors in pupillometry: (1) pupillary light reflex, (2) off-axis distortion, and (3) pupil dilation latency. For the issue of pupillary light reflex, all pupil dilation values of a participant were compared against the established individual baseline, calculating the absolute pupil increase in millimeters.[38] The pupil dilation values during pauses were compared against the separate baseline for the paused screen. Off-axis distortion occurs when the eye rotates away from the eye camera.[39] To compensate for this distortion, we used a geometric correction model developed by Hayes and Petrov.[40] For the pupil dilation latency, we removed the pupillometry data during the latency which was set as 800 ms from the onset of different stimuli (i.e., game screen and paused screen).[41]

After this preprocessing, we calculated means of the absolute pupil increase for each trial. To compare different periods (i.e. game-play and pause) in different modes (i.e., intense and plain), means for each period were computed. All pupil diameter samples with a value of 0 were discarded. Outliers lower than the first quartile – 1.5 interquartile range (IQR) and higher than the third quartile + 1.5 IQR were removed.

2.4.2. Performance Assessment

For performance assessment, we use the built-in scoring system of the AbcdeSIM game. The score calculates the number of correct decisions (e.g., applying correct oxygenation scores 20 points). Remaining time is included as additional points (e.g., higher score for more remaining time). Since the scoring system does not display whether a learner managed epilepsy in the SAB scenario, we extracted the corresponding information from game logs and computed the intensity-care score, by using Python and R.

3. Results

There were 4 missing records in eye-tracking data and 3 in game logs, due to technical issues such as crashes between the game system and the iView X software. The ratio of missing samples in eye-tracking data was 4.93%.

Table 5. Distributions of number of pauses taken per participant in PA

Pause number	Participant (%)
0	20(50.0)
1	8(20.0)
2	6(15.0)
3	5(12.5)
4	0(0.0)
5	1(2.5)
Sum	40(100)

Table 6. Descriptives of total scenario duration, total pause duration, and proportion of pause duration

	Min	Q_1	*Mdn*	Q_3	Max
Scenario duration (sec)	331	405	599	638	881
Pause duration (sec)	8	20	40	107	251
Proportion of pause duration (%)	1.4	3.6	6.7	17.8	44.3

Note. Q_1 = the first quartile, Mdn = median, Q_3 = the third quartile. Proportion of pause duration = Pause duration / Scenario duration.

3.1. Usage of Pauses

Table 5 shows the distribution of the total number of pauses taken by each participant in the PA condition. In average, 50.0% of the participants did not take any pauses at all, while the remaining 50.0% took pauses with total number ranging between 1 and 5. Table 6 shows the descriptives of total duration of scenarios and pauses.

3.2. Effects over Entire Procedure

Table 7 demonstrates the outcomes of all measures for cognitive load and performance. The outcomes were assorted between the different conditions of PU and PA, and between the subgroups of PA0 and PAn. PA showed higher cognitive load and performance than PU, while no consistent differences are shown between PA0 and PAn. There were no significant correlations between the measures except between game scores and intensity-care scores.

Table 7. Descriptives of measures for cognitive load and performance

Construct	Measure	Condition				Subgroup in PA			
		PU		PA		PA0		PAn	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
CL	Rating scale (0-100)	58.82	13.27	60.49	10.93	58.38	10.44	62.49	11.14
	Pupil increase (mm)	0.782	0.363	0.941	0.288	0.904	0.461	0.905	0.393
Performance	Game score	462.47	110.16	478.53	142.66	507.15	145.44	451.29	136.17
	Intensity-care score (%)	13.33	34.57	37.50	49.03	25.00	44.43	50.00	51.30

Note. PU = Pause unavailable, PA = Pause available, PA0 = trials with number of pauses taken is 0, PAn = trials with number of pauses taken is higher than 0.

4

The effects of allowing pauses on the different measures of cognitive load and performance are presented in Table 8. The existence of a positive effect of allowing pauses on cognitive load (H1a) is supported by pupil increase, while no significant effect is associated with the rating scale. To examine if this significant effect in pupil increase is still present even during the game-play periods, we excluded the pause periods from the pupil increase dataset in PA, then compared the data with PU again. The significant effect remained even after the exclusion. The existence of a positive effect of allowing pauses on performance (H1b) was supported by the intensity-care scores, while no significant effect was associated with game scores. In addition, scenario showed a significant effect with regard to pupil increase and game score, while displaying no significant effect on the rating scale. The level of interaction between the two fixed factors (i.e., condition and scenario) was low.
With regard to H2a and H2b, we tested the effect of taking pauses on the different measures of cognitive load and performance, by comparing PAn against PA0. No significant effects of actually taking pauses were found for any of the measures.

Table 8. Effects of allowing pauses and scenario difference on the measures of cognitive load and performance

Construct	Measure	Fixed effect	Estimate	Std.error	*t*(*z*) value	*P*
CL	Rating scale	(Intercept)	58.259	2.043	28.524	
		Condition (PA)	2.030	2.525	0.804	0.4214
		Scenario (SAB)	0.388	1.359	0.285	0.7753
	Pupil increase	(Intercept)	0.847	0.053	16.102	
		Condition (PA)	**0.126**	**0.064**	**1.970**	**0.0489**
		Scenario (SAB)	**-0.073**	**0.036**	**-2.013**	**0.0441**
(excluding pause periods)	Pupil increase	(Intercept)	0.848	0.054	15.763	
		Condition (PA)	**0.130**	**0.065**	**2.000**	**0.0455**
		Scenario (SAB)	**-0.073**	**0.037**	**-1.997**	**0.0458**
Performance	Game score	(Intercept)	499.870	20.610	24.257	
		Condition (PA)	14.120	24.400	0.578	0.5630
		Scenario (SAB)	**-69.190**	**17.220**	**-4.017**	**0.0001**
	Intensity-care score	(Intercept)	-1.872	0.539	-3.473	
		Condition (PA)	**1.361**	**0.629**	**2.163**	**0.0306**

Note. Significant effects (p <.05) are in boldface. Condition (PU) was compared as intercept for condition, while scenario (GIB) was compared as intercept for scenario.

Table 9. Pupil increase in different periods for different modes

Measure	Mode	Game-play		Pause		*t*	*df*	*p*	*d*	95%CI	
		Mean	*SD*	Mean	*SD*					lower	higher
Pupil increase (mm)	Plain	0.89	0.43	0.73	0.39	2.00	78	**0.049**	0.383	0.001	0.323
	Intense	1.38	0.42	0.46	0.53	4.90	10	**0.001**	2.091	0.503	1.335

Note. Significant effects ($p < .05$) are in boldface.

3.3. Effects for Pause Periods

Table 9 presents the descriptives of cognitive load during game-play periods and pause periods. These periods were also categorized into the two modes (i.e., plain and intense). The level of cognitive load was lower for the pause periods than the game-play periods in both two modes. The results from *t* tests showed a significant difference between pause periods and game-play periods for the plain mode with a small effect size ($t(78) = 2.00$, $p = .049$; $d = 0.38$) and for the intense mode with a large effect size ($t(10) = 4.90$, $p = .001$; $d = 2.09$). Figure 12 shows this effect on pause periods in different modes, in comparison with the cognitive load in different modes in PU.

Figure 12. Boxplot of pupil increase in different modes and conditions.

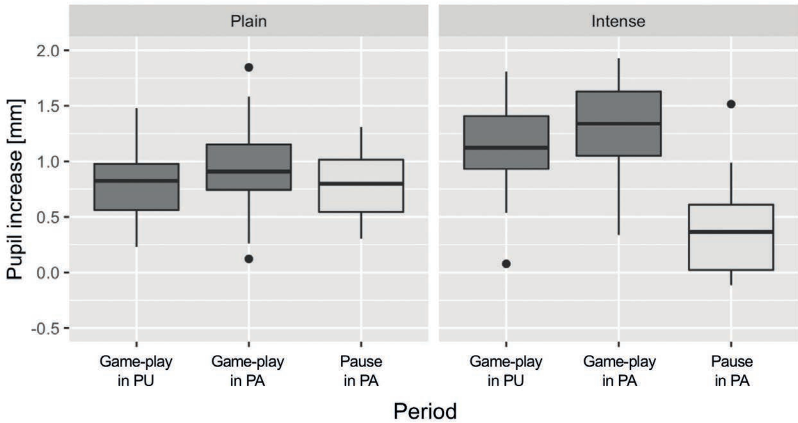


Figure 13. Distribution of difference in cognitive load between game-play periods and pause periods (i.e., pause – game-play).

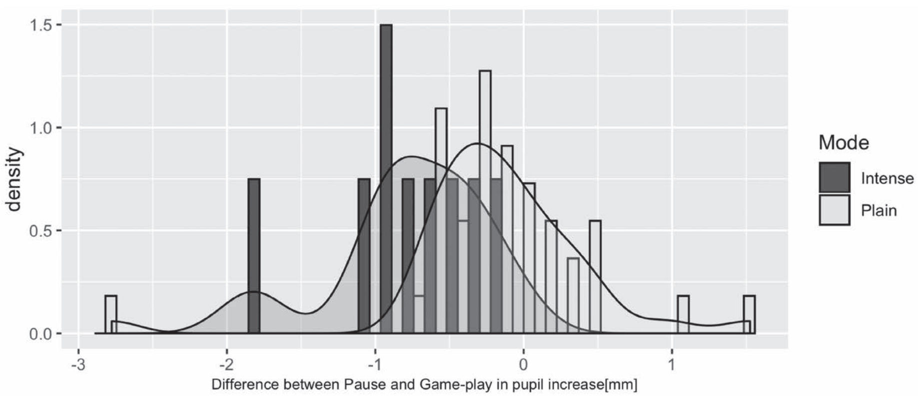
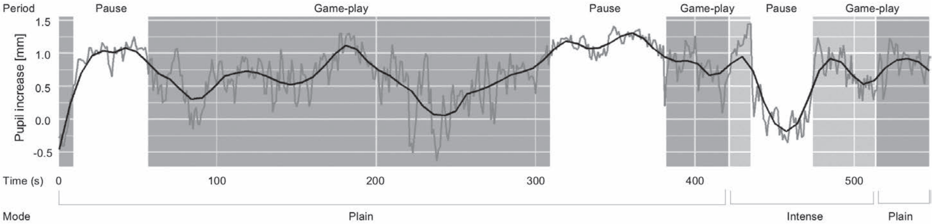


Figure 14. Time trend of pupil increase for one particular participant in PA.



Note. The annotation on the upper side indicates periods, while the annotation on the lower side indicates modes. A total of three pauses were taken: the first two pauses in plain mode, and the third pause in intense mode.

4

3.4. Exploration: Different Cognitive Activities during Pauses

H2c assumed the possible existence of different cognitive activities during pauses, namely, relaxation and reflection-on-action. To explore the diversity of these cognitive activities, we calculated the difference in cognitive load between game-play periods and pause periods, and plotted its distribution (Figure 13). The mean pupil increase during game-play periods was subtracted from the mean pupil increase in pause periods (i.e., pause – game-play) per trial. This difference was calculated for each mode separately, only when corresponding mean values existed for the trial. Although the distribution of differences was centered below zero (i.e., decrease of cognitive load during pauses) for both two modes, the differences were scattered widely (i.e., either decrease or increase of cognitive load during pauses), ranging from -2.78 to 1.53 (mean = -0.28, *SD* = 0.65). To demonstrate the different uses of pausing within individual participants, Figure 14 depicts a time trend of changes in pupil increase of one particular participant. This participant played SAH scenario in the PA condition, taking two pauses in plain mode and one pause in intense mode. As can be seen in the first two pauses in plain mode, the level of cognitive load increases, probably due to task-relevant cognitive activities. However, during the third pause in intense mode, cognitive load decreases, signifying that a relaxing pause took place.

4. Discussion

The present study tested the effects of pauses on cognitive load and performance in a medical simulation game. Two aspects of pausing were taken into account: availability of pauses (i.e., allowing pauses) and execution of pauses (i.e., taking pauses). The effects of these aspects of pausing on cognitive load and performance were examined at the overall level, while the cognitive load was further analyzed between pause periods and game-play periods. Pupillometry and a subjective rating scale were used for cognitive load measures, while game-score and intensity-care score were employed as performance measures. The main findings are as follows: (a) Allowing pauses increases cognitive load and performance at the overall level, regardless of whether pauses are actually taken; (b) Given that pausing is allowed, taking pauses gives no further benefit on cognitive load or performance at the overall level; (c) The level of cognitive load during pauses is lower than during game-play.

The first two hypotheses (H1a and H1b) pertained to the effects of allowing pauses. Focusing on this effect on cognitive load (H1a), a significant impact emerged via the pupillometry measure, while no effect was found for the subjective rating measure. We regard H1a as being supported agreeably by the data, not only because pupillometry is an objective measure that has been well-validated in the context of medical training environments, but also because the subjective rating might be a weak measure for certain forms of cognitive load in dynamic task-environments. The cognitive load caused by the availability of pauses mainly results from self-regulation (i.e., self-observation and self-control), which changes dynamically throughout performance. The subjective rating scale might not be an ideal measure to capture this form of cognitive load since it yields only one value for the entire performance phase. Moreover, supposing that this self-regulation was directly contributing to learning, the increase in cognitive load can be attributed to germane load. Considering that Young et al.[42] reported that germane load does not correlate with the subjective rating of perceived mental effort in a medical handover simulation, it is not surprising that the rating scale did not show a significant impact of allowing pauses in this study.

With regard to the effect on performance (H1b), the intensity-care score showed a significant impact, while no significant effect was found for game scores. We deem that H1b is at least partially supported, since the intensity-care score better represents certain aspects of performance in the context of self-regulation (i.e., self-observation and self-control) than the game score. In the SAB scenario, epileptic seizure attacks the patient randomly. A high intensity-care score can be achieved if the learner has been vigilant enough to counteract the sudden emergency, while playing the game systematically, which is highly relevant to self-observation and self-control. Therefore, we consider the intensity-care score to be more sensitive to changes in performance level due to the self-regulation, while game scores are less sensitive to these changes. Moreover, the built-in scoring system takes the form of a checklist of correct input, which has been reported as being a less effective measure of performance in a global manner.[43,44]

Given that H1a and H1b are at least partly supported, the results present a significant implication: allowing pauses (i.e., availability of taking pauses) is a distinct manifestation of the pause effect, that should be clearly distinguished from actually taking pauses (i.e., execution of pausing). Allowing pauses stimulates self-regulation in the performance phase (i.e., reflection-in-action), while taking pauses may stimulate self-regulation in the self-reflection phase (i.e., reflection-on-action). This difference emphasizes that the effect of allowing pauses is present throughout performance, regardless of whether pauses were actually taken or not. In fact, we found a robust effect of allowing pauses on cognitive load during performance, which remained even after excluding pause periods. This finding is also in line with Hasler et al.,[17] who found that learner-controlled pausing positively influences performance, despite the rare use of pause buttons.

Bridging self-regulated learning theory and cognitive load theory is a relatively new research line in recent years.[45] This study shows that combining the two perspectives provides a powerful framework for investigation of the pause effect. It robustly promotes construction of hypotheses and experimental design, leading to a stable explanation of the empirical results. However, this study did not reveal whether and how any self-regulation actually took place during performance. Future research could investigate the relevant cognitive activities using process-tracing methods such as think-aloud or retrospective reporting.[46]

The second category of hypotheses (i.e., H2a, H2b, and H2c) focused on the effects of actually taking pauses. Contrary to our expectations, the act of taking pauses did not show any significant impact on cognitive load nor performance at the overall level, failing to support H2a and H2b. We suppose that this is because the diverse activities during pauses were not controlled. In this study, we did not give any particular instructions about what to do during pauses, letting participants take pauses whenever they felt a need to do so. While this experimental design allowed us to explore learners' tendency to be involved in different cognitive activities during pauses, participants' cognitive activities seem to have failed to contribute to boosting beneficial cognitive processes (i.e., germane load) and performance at the overall level. We believe that, to make the act of taking pauses beneficial to learning and performance, providing guidance about the cognitive activities during pauses is crucial. For instance, students can be prompted to use pauses to recall relevant knowledge, evaluate the current diagnosis based on that knowledge, and establish a structured plan for the performance after the pause.

When comparing the cognitive load between different periods, the cognitive load during the pause periods was significantly lower than during the game-play periods, thus supporting H2c. It appears that the participants in this study have paused more often for relaxation than for reflection-on-action. This result opens up the potential of the use of relaxing pauses to manage the "helmet fire"[5] in medical simulation training. A high level of extraneous cognitive load caused by emotions heightened by intense situations have been reported as detrimental for medical simulation training with high-fidelity.[14,47] Considering that the effect size was larger in intense situations (i.e., intense mode) than in less-intense situations (i.e., plain mode), the use of relaxing pauses should provide substantial support for learners to manage the negative impact of high cognitive load.

Although relaxation has manifested itself as a predominant activity during pauses in this study, other activities (e.g., reflection-on-action) seem to be existing as well. As can be seen from the explorative investigation, some participants' cognitive load increases during pauses. Also, in-depth analysis of individual participants showed that one participant uses both reflective pauses and relaxing pauses, according to given situations (i.e., plain and intense modes). This diversity in cognitive activities is in accordance with the participants' responses to the questionnaire that we issued for exploratory purposes. When asked what they were doing during the pauses, the participants' responses varied. For instance, they were not only relaxing themselves, but also recapping their performance, overviewing possible options, recalling relevant knowledge, and planning what to do next.

While we adhered to experimental control for designated hypothesis-testing, we have not controlled specific pausing behaviors such as prescribed activities during pauses or duration of pauses. On the one hand, the exact activities that participants were engaged in during each pause were unknown. On the other hand, it allowed us to explore individual tendencies and to gain ecological validity, producing interesting results. This opens up multiple lines for future studies: What are the cognitive activities that take place during pauses? What makes these activities beneficial to performance and learning? And what guidelines for pausing should be provided to learners?

One notable implication of the results of this study is that the use of pupillometry as a cognitive load measure advances the investigation remarkably. Since pupil dilation is controlled by the autonomous nervous system, it reveals some aspects of cognitive load that subjective ratings cannot capture. Moreover, this continuous dataset enables window analysis, with which we can detect the net effect of allowing pauses by subtracting certain windows (i.e., pause periods). It also allows for measuring the temporary changes within specific periods, revealing the existence of different cognitive activities during these periods.

Interestingly, we have noticed that only 50% of participants in the PA condition took pauses, while the rest did not. Some researchers have reported similar observations of a rare use of pause buttons in educational animation, suggesting that students cannot locate an appropriate stopping point.[17,48,49] On top of that, we propose another explanation for the low usage of pausing; learners are not aware of the utility of taking pauses during performance. When the questionnaire survey asked the participants who did not pause at all about why they did not, some answered, "I thought pauses would not improve the given situation", or "Taking pauses seems to be unrealistic for real-life situations". Although time-outs are essential practices to reduce diagnostic errors and improve performance,[1] students might overlook the benefit of taking pauses during performance. Future research can study this phenomenon further, investigating why students hesitate to use pauses or whether encouraging them to use more pauses contributes to learning.

There are some limitations in this study. Firstly, this study used only two scenarios in a computerized simulation in the domain of emergency medicine. Testing these effects of pausing in other task-environments and in other domains is necessary. Secondly, it has not been tested whether the pause effects in simulation training can be replicated for the time-outs in real-task environments. Future study could investigate the possible differences between simulated-task environments and real-task environments. Thirdly, we focused on the pause effects on performance only, not learning or transfer to the workplace, in relation with cognitive load. Expanding the investigation to the effects on learning and transfer should be fruitful for medical education.

For clinical educators in simulation training, we suggest several approaches for the installment of pauses in training. Firstly, when employing the pause technique, the extra cognitive load that arises by simply allowing pauses should be taken into account. Although the extra processes (i.e., self-observation and self-control) are beneficial for learning and performance, it can also be detrimental if the total cognitive load exceeds the learner's working memory capacity. Thus, when prompting the use of pauses in a highly demanding simulation training, extraneous load from a high-fidelity environment can be lowered by adjusting the fidelity level (e.g., removing sound effects). Secondly, learners should be informed about the utility of pauses. While students can easily regard pausing as unhelpful and unrealistic, pausing is an essential practice that experts utilize, also known as the phenomenon of "slowing down when you should".[2] Practicing pausing in simulated task-environments can be a good opportunity to learn the "slowing down". Thirdly, learners should be taught how to utilize the two different types of pausing (i.e., reflective and relaxing) to enhance their performance and learning. Reflection can be encouraged to boost self-evaluation, while learners should manage their own cognitive load to prevent overload. When learners get panicked by intense situations, they could use pauses to take a deep breath and reorganize themselves.

This study has shown how students tend to use pauses without guidance, how the pause effect functions, and what should be the best use of pauses. The essential finding is that pausing is not a simple technique where cognitive load drops and performance enhances once learners take a pause. There are more complex layers of cognitive processes that do not manifest themselves at the physical level. Even in the case where a learner is not taking any pauses at all, which at first sight appears not to differ from the condition where pauses are not available, additional cognitive processes are already activated for the learner by merely considering about pauses. Also, although all acts of taking pauses appear to be similar at first glance, fundamentally different cognitive processes are running for each pause, depending on the given situation. To make pausing contribute to performance and learning, learners should be provided with some guidance to take these hidden cognitive processes into account. The best use of pauses is not a matter of availability of pauses nor the execution of pauses, but a matter of *how* to use pauses based on a deep understanding of *why* they should be taken.

References

1. Altpeter T, Luckhardt K, Lewis JN, Harken AH, Polk Jr HC. Expanded surgical time out: a key to real-time data collection and quality improvement. *J Am Coll Surg.* 2007;204(4):527-532. 10.1016/j.jamcollsurg.2007.01.009
2. Moulton CA, Regehr G, Lingard L, Merritt C, MacRae H. Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. *Acad Med.* 2010;85(10):1571-1577. 10.1097/ACM.0b013e3181f073dd
3. Trowbridge RL. Twelve tips for teaching avoidance of diagnostic errors. *Med Teach.* 2008;30(5): 496-500. 10.1080/01421590801965137
4. Sibbald M, De Bruin AB, Van Merrienboer JJ. Checklists improve experts' diagnostic decisions. *Med Educ.* 2013;47(3):301-308. 10.1111/medu.12080
5. Szulewski A, Egan R, Gegenfurtner A, et al. A new way to look at simulation-based assessment: the relationship between gaze-tracking and exam performance. *Canadian Journal of Emergency Medicine.* 2019;21(1):129-137. 10.1017/cem.2018.391
6. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001
7. Szulewski A, Gegenfurtner A, Howes DW, Sivilotti MLA, Van Merrienboer JJG. Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education.* 2017;22(4):951-968. 10.1007/s10459-016-9725-2
8. Sweller J. Cognitive Load during Problem-Solving - Effects on Learning. *Cogn Sci.* 1988;12(2): 257-285. DOI 10.1207/s15516709cog1202_4
9. Sandars J, Cleary TJ. Self-regulation theory: applications to medical education: AMEE Guide No. 58. *Med Teach.* 2011;33(11):875-886. 10.3109/0142159X.2011.595434
10. Young JQ, Sewell JL. Applying cognitive load theory to medical education: construct and measurement challenges. *Perspectives on Medical Education.* 2015;4(3):107-109. 10.1007/s40037-015-0193-9
11. Van Merrienboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ.* 2010;44(1):85-93. 10.1111/j.1365-2923.2009.03498.x
12. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* 2014; 36(5): 371-384. 10.3109/0142159X.2014.889290
13. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
14. Fraser K, Ma I, Teteris E, Baxter H, Wright B, McLaughlin K. Emotion, cognitive load and learning outcomes during simulation training. *Med Educ.* 2012;46(11):1055-1062. 10.1111/j.1365-2923.2012.04355.x
15. Leppink J, Van den Heuvel A. The evolution of cognitive load theory and its application to medical education. *Perspectives on Medical Education.* 2015;4(3):119-127. 10.1007/s40037-015-0192-x

16. Fraser KL, Ayres P, Sweller J. Cognitive Load Theory for the Design of Medical Simulations. *Simulation in Healthcare*. 2015;10(5):295-307. 10.1097/SIH.0000000000000097

17. Hasler BS, Kersten B, Sweller J. Learner control, cognitive load and instructional animation. *Appl Cogn Psychol*. 2007;21(6):713-729. 10.1002/acp.1345

18. Schwan S, Riempp R. The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and Instruction*. 2004;14(3):293-305. 10.1016/j.learninstruc.2004.06.005

19. Tabbers HK, de Koeijer B. Learner control in animated multimedia instructions. *Instructional Science*. 2010;38(5):441-453. 10.1007/s11251-009-9119-4

20. Lowe R. Interrogation of a dynamic visualization during learning. *Learning and Instruction*. 2004; 14(3):257-274. 10.1016/j.learninstruc.2004.06.003

21. Lowe R. Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education*. 1999;14(2):225-244. 10.1007/Bf03172967

22. Schnotz W, Bockheler J, Grzondziel H. Individual and co-operative learning with interactive animated pictures. *European Journal of Psychology of Education*. 1999;14(2):245-265. Doi 10.1007/Bf03172968

23. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. *Educ Psychol*. 2003;38(1):43-52. 10.1207/S15326985ep3801_6

24. Mayer RE, Chandler P. When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *J Educ Psychol*. 2001;93(2):390. 10.1037/0022-0663.93.2.390

25. Moreno R. Optimising learning from animations by minimising cognitive load: Cognitive and affective consequences of signalling and segmentation methods. *Appl Cogn Psychol*. 2007; 21(6):765-781. 10.1002/acp.1348

26. Spanjers IAE, Van Gog T, Wouters P, Van Merrienboer JJG. Explaining the segmentation effect in learning from animations: The role of pausing and temporal cueing. *Computers & Education*. 2012;59(2):274-280. 10.1016/j.compedu.2011.12.024

27. Zimmerman BJ. Becoming a self-regulated learner: An overview. *Theory into Practice*. 2002; 41(2):64-70. DOI 10.1207/s15430421tip4102_2

28. Seufert T. The interplay between self-regulation in learning and cognitive load. *Educational Research Review*. 2018;24:116-129. 10.1016/j.edurev.2018.03.004

29. Winne P, Hadwin A. Studying as self-regulated engagement in learning. in metacognition in educational theory and practice. *Metacognition in Educational Theory and Practice*. 1998;93: 27-30.

30. Szulewski A, Roth N, Howes D. 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. *Acad Med*. 2015;90(7):981-987. 10.1097/Acm.0000000000000677

31. Aldekhyl S, Cavalcanti RB, Naismith LM. Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on Medical Education*. 2018;7(1):23-32. 10.1007/s40037-017-0392-7

32. Korbach A, Brünken R, Park B. Differentiating different types of cognitive load: A comparison of different measures. *Educ Psychol Rev*. 2018;30(2):1-27. 10.1007/s10648-017-9404-8

33. *AbcdeSIM* [computer program]. Rotterdam: VirtualMedSchool; 2012.

34. Paas FGWC. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics - a Cognitive-Load Approach. *J Educ Psychol.* 1992;84(4):429-434. 10.1037/0022-0663.84.4.429

35. Holmqvist K, Andersson R. *Eye-tracking: A comprehensive guide to methods, paradigms and measures.* Lund, Sweden: Lund Eye-Tracking Research Institute; 2017.

36. Bates D, Machler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software.* 2015;67(1):1-48. 10.18637/jss.v067.i01.

37. *R: A language and environment for statistical computing* [computer program]. R Foundation for Statistical Computing; 2019.

38. Beatty J, Lucero-Wagoner B. The pupillary system. In: Cacioppo JT, Berntson G, Tassinary LG, eds. *Handbook of Psychophysiology.* Vol 2. Cambridge: Cambridge University Press; 2000: 142-162.

39. Mathur A, Gehrmann J, Atchison DA. Pupil shape as viewed along the horizontal visual field. *Journal of Vision.* 2013;13(6):3-3. 10.1167/13.6.3

40. Hayes TR, Petrov AA. Mapping and correcting the influence of gaze position on pupil size measurements. *Behav Res Methods.* 2016;48(2):510-527. 10.3758/s13428-015-0588-x

41. Ahern S, Beatty J. Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science.* 1979;205(4412):1289-1292. 10.1126/science.472746

42. Young JQ, Irby DM, Barilla-LaBarca ML, Ten Cate O, O'Sullivan PS. Measuring cognitive load: mixed results from a handover simulation for medical students. *Perspectives on Medical Education.* 2016;5(1):24-32. 10.1007/s40037-015-0240-6

43. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education.* 1996;1(3):227-233. 10.1007/BF00162920

44. Dankbaar ME, Stegers-Jager KM, Baarveld F, et al. Assessing the assessment in emergency care training. *PLoS One.* 2014;9(12). 10.1371/journal.pone.0114663

45. Sweller J, Van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev.* 2019;31(2):261-292. 10.1007/s10648-019-09465-5

46. Van Gog T, Paas F, Van Merrienboer JJ, Witte P. Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *J Exp Psychol Appl.* 2005; 11(4):237-244. 10.1037/1076-898X.11.4.237

47. Fraser K, Huffman J, Ma I, et al. The emotional and cognitive impact of unexpected simulated patient death: a randomized controlled trial. *Chest.* 2014;145(5):958-963. 10.1378/chest.13-0987

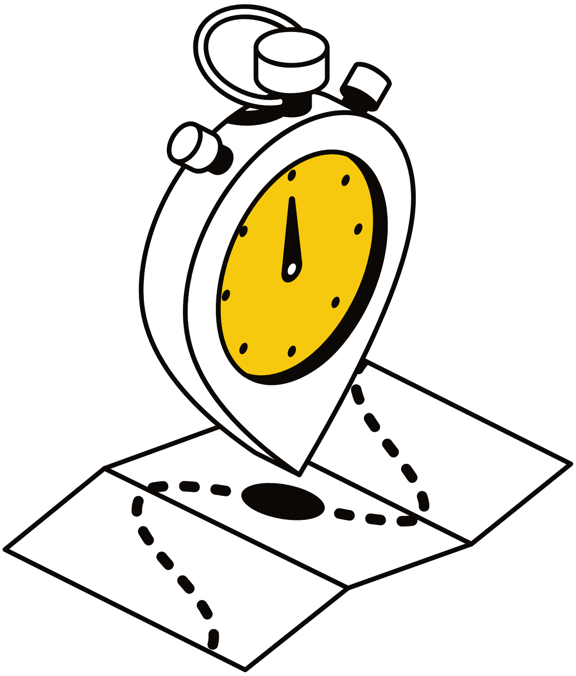
48. Kettanurak V, Ramamurthy K, Haseman WD. User attitude as a mediator of learning performance improvement in an interactive multimedia environment: an empirical investigation of the degree of interactivity and learning styles. *Int J Hum Comput Stud.* 2001;54(4):541-583. 10.1006/ijhc.2001.0457

49. Moreno R, Valdez A. Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Etr&D-Educational Technology Research and Development.* 2005;53(3):35-45. Doi 10.1007/Bf02504796

Chapter 5

The Reflective Pause in Simulation Training: The Effects of In-Action Reflection on Cognitive Load and Performance

This chapter is under revision as:
Lee JY, Donkers J, Jarodzka H, Sellenraad G, Faber TJE, Van Merriënboer JJG. The reflective pause in simulation training: The effects of in-action reflection on cognitive load and performance.



Abstract

The reflective pause, taking a pause during performance to reflect, is an important practice in professional training that fosters learning processes. However, research on the reflective pause has been scarce, and consensus on effective reflection-prompting methods is lacking. Using educational theories, we propose how to design cognitive and metacognitive aids to support for reflective pause in a computer-based simulation task environment. Assuming this reflective pause with the aids helps to optimize cognitive load and allows for restructuring of mental models, we examine its effects on four aspects of performance and learning: cognitive load, domain-specific and domain-general aspects of performance, and the structure of cognitive schemas. Medical students ($N = 72$) performed tasks in a computerized simulation training for emergency medicine, in two conditions: reflection condition ($n = 36$) where reflection was prompted during pauses, and control condition ($n = 36$) without such prompts. The effects of reflective pauses emerged in the later stage of the learning process, while no significant effects were identified in the earlier stage. Cognitive load decreased and domain-general aspects of performance improved. However, domain-specific performance aspects and schema structure did not, probably due to lacking feedback during reflection. These results suggest that theory-based support design can make in-action reflection effective, demonstrating that reflective pauses can enhance performance, but an adaptation period is required.

Keywords: Games, human-computer interface, post-secondary education, Simulations, 21st century abilities

1. Introduction

In critical fields, such as firefighting, aviation, surgery, and emergency medicine, the tasks are extremely dynamic and hazardous.[1] During performance, practitioners should constantly adapt their actions, which are in turn associated with finality and significant consequences (e.g., death of patient or team members).[2,3] This gives rise to two important perspectives on training for these professions: simulation-based training and adaptive performance.[3] Through practice in simulated scenarios, practitioners must learn to adapt their performance by making a series of urgent yet careful decisions. To make these decisions successful, the *reflective pause*,[4] stopping the current course of action for a brief period of time to reflect, becomes an essential learning strategy for the critical domains.[2,3] In medical education, researchers have called for more attention to this in-action reflection, as opposed to post-action reflection (e.g., debriefing), as medical interventions become more dynamic and fast-paced and the standards for patient safety increase.[5-7] However, the reflective pause has not yet been fully acknowledged in education and workplace learning. Students are often unaware of the utility of taking pauses.[8,9] Educators have not sufficiently considered the reflective pause as part of their curricula, only focusing on algorithmic task approaches that hamper critical thinking.[10,11] Health professionals are often unwilling to interrupt current performance by pausing since they assume it to be inefficient.[12,13]

In education, it has been already reported that students tend to disregard in-action reflection during task performance. Students think prompts of in-action reflection disturbing and focus on problem-solving itself, ignoring the prompt either unknowingly or purposely.[14,15] Moreover, educational support for in-action reflection is lacking,[16,17] and much debate about how reflection should be prompted is still ongoing.[18-20] Studies on the effects of reflective pause have reported mixed results because different setups of when and how to reflect significantly diverge the effects.[21] Researchers have pointed out that lack of understanding of this diversity can hinder the development of research,[21,22] and that real-time measures (e.g., eye-tracking, screen recordings, physiological sensors) should be leveraged to track and understand the effects of in-action reflection during performance.[21,23,24]

This study contributes to the literature with three added values. First, we propose that, to make in-action reflection effective, the when and how to reflect should be designed a priori based on task analysis and educational theories. Most research on in-action reflection in computer-based task environments have used system-driven prompts, which lacks depth and details comparing to instructor-driven post-action prompts.[20,22] Our study presents an example of this design using a computer-based task environment, to empirically examine its effects. Second, while many studies on reflection focus on its

effects on general learning outcomes or self-regulative behaviors,[21,25,26] we systematically categorize different aspects of performance and learning that can be affected by the reflective pause, by using the frameworks of cognitive load theory and the medical pause.[7,27] Third, on top of using self-reported data, we employ real-time objective measures such as eye-tracking and game logs in a computer-based simulation, to provide in-depth analysis of the effects of in-action reflection on performance.
In the following sections, we first conceptualize how the reflective pause affects different aspects of performance and learning. We then describe what cognitive and metacognitive aids could be provided to make the training of reflective pausing effective. Lastly, we present a set of hypotheses about the effects of the reflective pause, which will be tested empirically.

1.1. The Effects of the Reflective Pause

Cognitive load theory (CLT) is a powerful framework to explain learning processes in simulation training.[27-29] It posits dynamics of diverse cognitive processes in simulation that impose working memory that has limited capacity, which allows for an explanation of the changes in performance real-time. Following the typology suggested by Lee et al.,[7] three types of cognitive load can be conceptualized in simulation training: primary load (PL) imposed by domain-specific task performance (e.g., a resuscitation task), secondary load (SL) from domain-general processes that assist primary task performance (e.g., vigilance on situational changes during the resuscitation), and extraneous load (EL) from processes irrelevant to the task performance (e.g., being distracted by noise from a computer-based simulation).

Within this framework, the effects of the reflective pause on both learning and performance can be described further. CLT presumes that performance deteriorates when the sum of the three types of loads either overloads or underloads working memory. Coined *negative momentum*, this suboptimal status can be interrupted by means of taking a pause.[7] During this pause, working memory is reinstated by restructuring mental models (i.e., schemas) in long-term memory, which in turn promotes learning. For example, in an intense resuscitation scenario, practitioners may experience cognitive overload losing their attention to patients' safety. They then take a pause to recap the current situation. By connecting information elements from working memory (e.g., previous actions) and long-term memory (e.g., alternative solutions), schemas are reconstructed (e.g., planning next priorities) and working memory is reinstated.[27,30] As a result, cognitive load is optimized and subsequent performance improves (e.g., improvement in diagnosis and attention to patient safety).

For critical fields, the reflective pause is of utmost importance for three reasons. First, many adverse events are caused by human error (e.g., misdiagnosing patients) in these fields.[31,32] As tasks in the fields have become even more dynamic and complex in recent years, practitioners are required to be extra circumspect and mindful.[2,6,33] Second, the tasks in extreme environments often feature high levels of uncertainty where important information emerges during task execution.[2] When the new information comes up, practitioners are required to pause and reflect in order to redirect their task performance.[3] Third, the reflective pause can facilitate learning within the context of performance.[7,34] Since the tasks in critical domains are highly context-dependent and lengthy (e.g., a surgery lasting several hours), creating learning opportunities during the performance is crucial.

1.2. Cognitive and Metacognitive Aids (CMAs)

Despite its benefits, taking a reflective pause is not an easy skill to acquire for novices. Aside from the negative momentum, there is the positive momentum of workflow that should not be interrupted, otherwise it can also impair performance.[35] Thus, the decision making on pausing while balancing between the negative and positive momentum is a highly complex skill that includes clinical judgement and time management.[7,36] Moreover, it is a metacognitive process to monitor and regulate one's own performance (i.e., self-regulation),[37] requiring additional support for this aspect.[38,39]

To train self-regulatory skills for novices, *second-order scaffolding* is necessary[40]: At the beginning of training of self-regulatory skills, educators should hold maximal control over important decision-making and provide full levels of support, while this control level decreases later as the learner becomes more experienced. In education, first-order scaffolding applies to domain-specific skill training, while second-order scaffolding pertains to training of domain-general skills such as self-regulation.[40] According to instructional design theories, second-order scaffolding for reflective pausing can be established by providing *cognitive and metacognitive aids* (CMAs) including prompts, cues, and leading questions.[40-42]

1.2.1. Prompts

When to pause can be proactively planned and initiated by educators in training programs.[5,7] To locate good pause timings, educators should first assess the learners' competence and prompt pauses within task partitions that are meaningful for the learner. According to CLT, logically segmenting learning content into smaller chunks and inserting pauses between these segments may facilitate the comprehension of the content structure.[43,44] In complex domains such as medicine, creating skill clusters for different task phases and arranging their sequence is necessary for skill training, so called, *part-task sequencing* methods.[40] For example, *forward chaining with snowballing* is a useful part-task sequencing method for teaching medical interventions. It arranges the skills necessary for task performance in a natural order as they are applied in practice. While simple forward chaining arranges the parts one by one (e.g., part A – part B – part C), the snowballing method includes the previous clusters in each new cluster (e.g., part A – part A and B – part A, B, and C).[40] As this method is designed to maximize the interconnection between different skill clusters, implementing reflective pauses between these clusters to contemplate this connection would make the learning processes even more effective.

1.2.2. Cues

Once the timing of a pause is prompted properly, the question of what to reflect on arises. Although reflection is a high-level cognitive process that benefits learning, a well-known pitfall is its dependency on episodic memories, which often leads to forgetting or fabrication.[45,46] To compensate for this drawback, the reflection process can be supported by cues, records of learners' performance that cue their memories of the performance.[47] For computer-based simulation task environments, combining video recording and eye movements can create promising cues, as eye movements reflect cognitive strategies to collect visual information on computer screens.[48-50] Studies have shown that reflection with these cues becomes longer in duration, includes more information on metacognitive processes, and is more detailed in visual strategy description than reflection without cues.[48,51,52]

1.2.3. Leading Questions

Despite the availability of effective cues, reflection can still be ineffective if the learner does not know *how* to reflect. Since learners often lack well-organized schemas in long-term memory, they do not apprehend what to discover through reflection, concentrate on surface-level features from the cues, and lose learning opportunities to restructure their schemas.[53,54] To guide them with what to discover from the cues, leading questions can be provided, in a so-called *guided discovery method*.[55] These questions help learners start from what they already know and develop their schemas by self-explaining the relationships between pieces of information.[42,56] This method includes generating hypothetical cases, selecting counterexamples, and entrapping learners. For instance, learners in medical domains can be asked to rearrange their diagnosis and predict its effects, or to present a counterexample of their diagnosis that would lead to an undesirable consequence.[40]

1.3. Present Study

This study empirically examines the effects of the reflective pause supported by CMAs on different aspects of performance and learning in simulation training. Given the benefits of the CMAs discussed above, the reflective pause is expected to positively affect learning processes. We use a computer-based simulation where an intense resuscitation scenario in emergency medicine is presented. As this scenario provokes high levels of cognitive load for students,[57] we expect that, in a condition where the reflective pause is applied (RP), cognitive load would be optimized by decreasing it, as students reinstate their working memory through reflection (H1).

Thanks to this reinstatement and related schema restructuring, other aspects of performance and learning are expected to improve. First, the domain-specific primary performance (e.g., diagnosis and intervention) is enhanced by interconnecting with schemas in long-term memory (H2). Also, the domain-general secondary performance (e.g., vigilance on patient status) improves, since the optimization creates more space for the secondary load (H3). As a result of the interconnection with long-term memory, learning is promoted, thus cognitive schemas improve (H4). These hypotheses are recapitulated as below:

H1. RP decreases cognitive load.
H2. RP improves primary performance (i.e., diagnosis and intervention).
H3. RP improves secondary performance (i.e., vigilance on patient status).
H4. RP furthers development of cognitive schemas.

The corresponding measures for each hypothesis are described in Table 10, of which the details of operationalization will be elaborated in the next section.

Table 10. Hypotheses, constructs, and measures

Hypothesis	Construct	Measure
H1	Cognitive load	Pupil increase (mm) Paas scale (1-9)
H2	Primary performance (Diagnosis and intervention)	Game score
H3	Secondary performance (Vigilance on patient status)	Transition rate across the vital signs monitor (VSM) (/min)
H4	Schema structure	Handover report

2. Material and Methods

2.1. Participants and Design

A total of 76 medical students (53 females; mean age = 22.3; *SD* = 2.4) was recruited from Maastricht University, the Netherlands. The participants' academic years varied from 2^{nd} to 6^{th}. They all had taken the basic course of emergency medicine in their 1^{st} year, while their experience in the domain varied according to their individual curriculum. Using a between-subjects design, we randomly assigned participants into two groups:

the experimental group where participants were prompted to reflect on their performance during pauses (RP condition), and the control group where they were asked to perform an irrelevant task instead during the same pauses (e.g., assessing random simulation games) (No-RP condition). For all participants, we provided three pauses between part-tasks grouped by the skill clusters. Their performance in each cluster was repeatedly measured for each participant via the measures described in Table 10.

2.2. Materials and Apparatus

2.2.1. A Medical Simulation Game

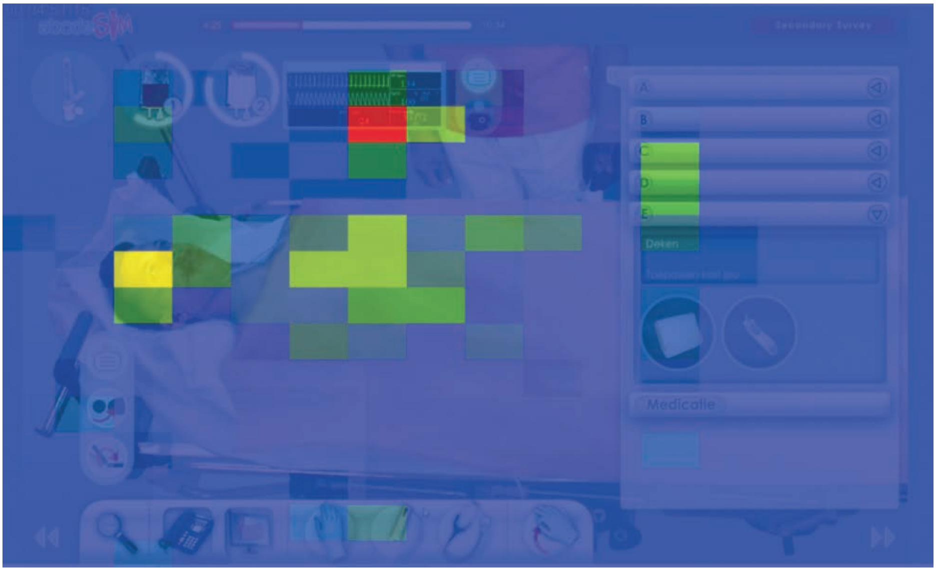
We used a computer-based simulation game for emergency medicine, AbcdeSIM.[58] The five letters ABCDE stand for the five phases that should be followed for acute care (i.e., Airway, Breathing, Circulation, Disabilities, and Exposure). This game simulates real-life situations in the emergency room, where a high-fidelity human physiology is programmed to react to the learners' interaction. The gastrointestinal bleeding (GIB) scenario was used since its level of complexity is high enough to provoke salient cognitive load. This scenario presents a 32-year-old male patient with hypovolemic shock. The primary task is to stabilize this patient by diagnosing (e.g., laboratory examination) and intervening (e.g., fluid infusion). As this patient's vital signs change acutely and dynamically, extra attention to the vital signs monitor (VSM) is required.

The game was run on a personal computer (Intel Core i7 2.67 GHz CPU, 1.98 GB RAM), and presented on an LCD monitor (1650 x 1080) and a headset for sound effects. Participants interacted with the game using a wireless mouse. The game score was calculated by summation of correct in-game actions of diagnosis and intervention, yielding the measure to test H2 (i.e., the primary performance).

2.2.2. Cognitive and Metacognitive Aids (CMAs)

To control the pause prompts for the two conditions, the same number of pauses was inserted between task clusters. These task clusters were designed using the forward chaining with snowballing method. As the task of the AbcdeSIM game contains the five phases that should be performed sequentially, we translated the phases into five skill clusters necessary for Airway (A), Breathing (B), Circulation (C), Disabilities (D), and Exposure (E). These skill clusters were then composed into three task clusters: AB – ABC – ABCDE. A current physician in emergency medicine designed this clustering and time-on-task required for each cluster, so that it fit the learning processes for the GIB scenario.

Figure 15. Heatmap as a type of cognitive and metacognitive aid



The cues to support reflection included a gameplay recording where eye movements are superimposed on the gameplay screen, and a heatmap that shows eye fixation allocation (Figure 15). The former is expected to show the learner's performance and attention allocation in a timeline, while the latter shows the overall attention allocation in summary. The game play and eye movements were recorded and displayed by SMI BeGaze software (version 3.6, www.smivision.com).

Using the guided discovery method, a set of leading questions was designed. These questions had two categories of diagnosis and intervention where each category included two questions of what should have been done and what should have not been done, resulting in four questions: "Did I miss something important to check?", "What did I check unnecessarily?", "Did I miss any important intervention?", and "What redundant interventions have I applied?" Lastly, an overall question, "How to improve my performance?", was added, resulting in a total of five questions. The participants were asked to verbalize their reflection (i.e., think-aloud). This verbalization was recorded, and the number of words was counted for an additional analysis on the reflection.

2.2.3. Eye-Tracking

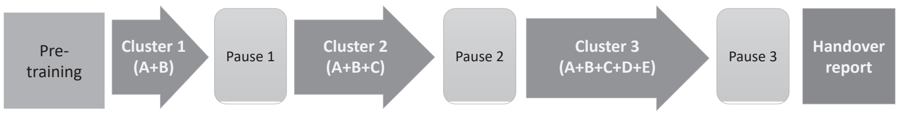
Eye-tracking data was collected by an SMI RED remote eye-tracker with a sampling rate of 250Hz and SMI iView X software (version 2.7.13). We used a dual-computer setup where a laptop computer with the iView X software was connected to a personal computer for game presentation. We used SMI Experiment Center 3.5 software (version 3.2.11) to arrange calibration and presentation of instruction and stimuli. A forehead-and-chin rest was used to prevent head movements and make the geometry between the eye and the eye tracker stable.[59] The experiment room maintained a constant luminance, without windows or noise, composing a controlled environment for eye-tracking.

2.2.4. Cognitive Load Rating and Handover Report

We used the Paas scale,[60] a subjective rating scale for cognitive load, for triangulation. This triangulation is expected to allow for a more comprehensive perspective to the results on cognitive load than using pupillometry only,[61] The Paas scale has 9-point ratings with the value 1 representing the lowest cognitive load and the value 9 representing the highest.

The handover report is a brief document where the patient case is communicated in an encapsulated way for clinical handover. It requires the development of structured schemas in clinical cases,[62] thus, it can provide a reliable measure to test H4 (i.e., schema structure construct). Two current physicians assessed the participants' reports, based on the SBAR,[63] the internationally well recognized protocol for clinical handover. Their scores were averaged for data analysis, and the inter-rater reliability was examined (r (69) = 0.98, p < .001).

Figure 16. Procedure of a session



2.3. Procedure

Figure 16 shows the entire procedure of a session. The participants were individually invited to the eye-tracking laboratory. They signed informed consent and filled out a demographic questionnaire. In the pre-training, they watched a tutorial about the game-play, and played an easy scenario to familiarize themselves with the game functions. Participants in RP were additionally instructed in how to use the CMAs (i.e., cues and leading questions) for their reflective pauses. This instruction included tutorials of how to view and interpret the cues (i.e., game replay with eye movements, and heatmap) by using the SMI BeGaze software, and examples of application of the leading questions to reflection. All participants, then, were positioned for the eye-tracking setup with a chin rest, and followed a 9-point calibration procedure. A scrambled image of the game screen was presented for 5 seconds to establish a baseline for pupillometry. Next, the participants played the first task cluster (i.e., AB) for 2 minutes, with a timer visible on the screen. When the time was up, the first pause was prompted. During this pause, participants in RP reflected on their performance using the cues and the leading questions. They were asked to think-aloud their reflections, and this think-aloud was recorded through a microphone. During the same pause, the participants in No-RP watched an advertisement video of an unrelated medical simulation game. They then were asked to assess this game in the advertisement in writing.

For the second (i.e., ABC) and the third (i.e., ABCDE) task cluster, the same procedure from the calibration till the activities during a pause was repeated. The duration of gameplay was 4 and 6 minutes for the second and third cluster, respectively. After each cluster, the participants rated their perceived cognitive load using the Paas scale. Lastly, after the third pause, the participants filled in the handover report, being asked to make it as concise and comprehensive as possible. In addition, a short questionnaire for perceived utility of the CMAs ensued. The entire session took about 40 minutes.

2.4. Data Analysis

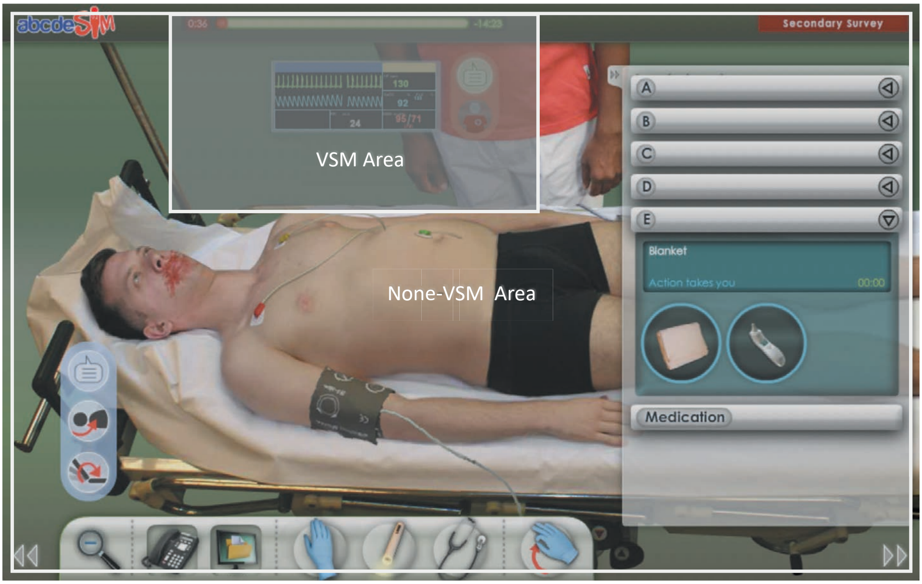
2.4.1. Eye-Tracking Data

The eye-tracking data included sampled measurements of pupil diameter and eye fixation location. This data was merged with game logs by synchronizing the game system with the eye-tracking software, then imported into R version 3.5.1,[64] for data processing. Pupillometry is well-known as an objective physiological measure of cognitive load in medical simulation settings.[29,65] To test H1 (i.e., the cognitive load construct), we used the pupil diameter data to measure real-time cognitive load. All pupil diameter samples with a value of 0 were removed. We then controlled for confounding factors in pupillometry, namely, pupillary light reflex, off-axis distortion, and pupil dilation latency (see Lee et al.[8] for details). As a result, means of absolute pupil increase against the baseline were calculated for each cluster.

The fixation location data was processed to produce transition rates. Transition rate refers to the number of gaze shifts per second from one area of interest (AOI) to another, which can represent cognitive processes to collect information across different AOIs.[59] We defined the upper-middle area of the screen where the VSM is located as an AOI (Figure 17). The transition rate between this VSM AOI and none-VSM areas was expected to reflect the vigilance on patients' status, resulting in the test of H3 (i.e., the secondary performance). Since some pop-up windows can occasionally block the

VSM areas (e.g., conversation with the virtual nurse), we extracted pop-up events from the game logs and modified the AOIs accordingly, by using Python (version 3.9.6).[66]

Figure 17. AOI definition: vital signs monitor (VSM) area versus none VSM area



2.4.2. Statistical Analysis

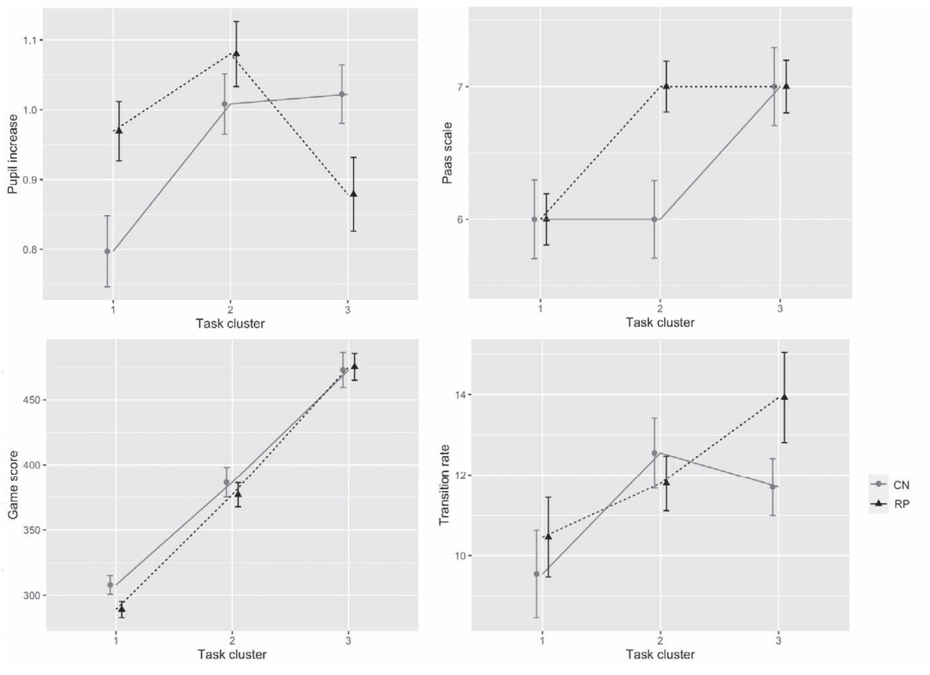
Outliers were removed using the Tukey method [67]. To investigate the effects of reflective pauses across all clusters, we first fitted linear mixed-effects models, using the lme4 package[68] in R. "Condition" and "Cluster" were entered as fixed effects with interaction term, while "Participant" was treated as a random factor (random intercept). To see whether the effects exist in specific clusters, we built separate models with dummy variables. For the effects of the first pause on the second cluster, we defined dummy variables for Cluster correct for the difference at the first cluster. For the effects of the second pause on the third cluster, another dummy encoding for Cluster was used to correct for the difference at the second cluster. For all models, residual plots were visually inspected and did not reveal any obvious deviations from homoscedasticity or normality. Likelihood ratio tests of the full model with the effects in question against the model without these effects obtained p values.

To calculate correlations between the reflection length (i.e., word count of the reflection during pauses) and performance measures, we used the Kendall's method. R was used for all statistical analysis. We considered $p < .05$ to be statistically significant, except for the effects in specific clusters where $p < .025$ was applied to correct for the multiple comparisons.

3. Results

Six clusters in the eye-tracking data and one cluster in the game logs were missing due to technical issues in the connection between the game system and the iViewX software. For the data quality of transition rate, eye-tracking data with tracking ratio below 50% and accuracy deviation larger than two degrees were excluded. After the exclusion, the average tracking ratio was 96.9%, and the average accuracy deviation was 0.9 degrees.

Figure 18. Performance change over task clusters



Note. Significant effects are marked with asterisks, which existed in two measures: pupil increase and transition rate. These effects emerged only in the third cluster with the reference to the second cluster.

Table 11. Descriptive of performance over task clusters.

Construct	Measure	No-RP						RP					
		Cluster1		Cluster2		Cluster3		Cluster1		Cluster2		Cluster3	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Cognitive load	Pupil increase (mm)	0.80	0.29	1.01	0.25	1.02	0.24	0.97	0.24	1.08	0.26	0.84	0.34
	Paas scale [a] (1-9)	6	1.75	6	1.73	7	1.77	6	1.15	7	1.13	7	1.19
Primary performance	Game score	308	41.8	387	67	473	77.2	289	36.1	377	54.8	475	60
Secondary performance	Transition rate (/min)	9.55	6.4	12.5	5.14	11.7	4.26	10.5	5.83	11.8	3.9	14.6	6.47
Schema structure	Handover report	5.4 (2.81)						5.64 (1.59)					

[a] Median was used instead of mean as a central tendency.

3.1. Effects of Reflective Pauses

Table 11 shows the descriptives of all measures of performance. The outcomes were sorted by the three clusters. In general, pupil increase was enlarged over the clusters, except that it decreased at the third cluster in RP, indicating lower cognitive load. The Paas scale resulted in high scores overall (i.e., between 6 and 7) regarding the 9-point ratings. Game score increased throughout all clusters. Transition rate also increased over the clusters, except that it decreased at the third cluster in No-RP, indicating lower vigilance on patient status. Handover report score was somewhat higher in RP. Figure 18 visualizes these changes in the measures. Only the measures that are applicable to the clustering were included in this visualization.

To examine statistical significance in the effects of reflective pauses, we used linear mixed-effects models. When comparing RP against No-RP in general, the models did not show any significant difference between the two conditions in any of the measures. However, when looking at the separate clusters, the third cluster showed a significant difference in two measures: pupil increase and transition rate. Table 12 presents the effects of Condition at the second and the third clusters on the different measures.

Table 12. Effects of reflective pauses on the measures of cognitive load and performance.

Construct	Measure	Effect	β	*SE*	*t* value	*p*
CL	Pupil increase	Condition(RP) in Cluster 2	-0.077	0.066	-1.160	0.248
		Condition(RP) in Cluster 3	-0.245	0.066	-3.717	**0.000**
	Paas scale	Condition(RP) in Cluster 2	0.027	0.254	0.107	0.915
		Condition(RP) in Cluster 3	0.011	0.254	0.045	0.964
Primary performance	Game score	Condition(RP) in Cluster 2	11.760	15.930	0.738	0.462
		Condition(RP) in Cluster 3	-11.759	15.926	-0.738	0.462
Secondary performance	Transition rate	Condition(RP) in Cluster 2	-1.717	1.826	-0.941	0.349
		Condition(RP) in Cluster 3	4.254	1.850	2.299	**0.024**
Schema structure	Handover report	Condition(RP)	-0.154	0.613	-0.252	0.802

Note. Significant effects ($p < .05$) are in boldface. Cluster 1 was the reference for the Cluster 2 effect, and Cluster 2 was the reference for the Cluster 3 effect.

To test H1 (RP decreases cognitive load), we used pupil increase and the Paas scale. Pupil increase showed a significant difference between the conditions at the third cluster, while the Paas scale did not show any significant difference. For H2 (RP improves primary performance), the game score did not show any significant effects. The measure for H3 (RP improves secondary performance) demonstrated a significant effect in the expected direction, again at the third cluster. No significant effects were detected in the handover report measure for H4 (RP furthers development of cognitive schemas), although the difference was in the expected direction. The significant effects are marked with asterisks in Figure 18.

In addition, to see whether any difference exists in the reference (i.e., baseline), a *t* test was used for each measure in the first cluster. The results of this test are expected to show whether the pre-training influenced the baseline of each measure. There was a significant difference of the condition only in pupil increase: It was higher in RP ($M = 0.97$, $SD = 0.24$) than in No-RP ($M = 0.80$, $SD = 0.29$) ($t\,(63) = -2.59$, $p = .012$, $d = 0.64$, 95% CI [-0.30, -0.04]).

Table 13. Correlations between reflection length measured by word count and performance measures.

Reflection	Pupil Increase	Paas scale	Game score	Transition rate	Handover report
Reflection connected to the previous cluster	0.06 [-.07, .19]	**0.27** [.14, .40]	**0.37** [.25, .49]	**0.15** [-.03, .32]	NA
Reflection connected to the next cluster	0.02 [-.12, .19]	**0.31** [.13, .49]	**0.24** [.08, .39]	0.03 [-.17, .23]	-0.07 [-.37, .22]

3.2. Exploration: Reflection Length and the Utility of the CMAs

As an exploration, we looked closer into the reflection activity in RP. First, we examined whether the length of reflection is related to the performance in the previous or following clusters. Table 13 shows the correlations between reflection length measured by word count and performance measures. When compared with the previous cluster, the length of reflection was positively correlated with the scores on the Paas scale, game score, and transition rate. When pairing with the next cluster, the correlation existed only with scores on the Paas scale and game score.

We then explored how students subjectively perceived the reflection activity. When asked how helpful the reflection activity was for performance and learning, participants rated it as 75%, on a scale where 0% represents not helpful at all and 100% represents fully helpful. We also asked them why it was helpful or why it was not, via a multiple-choice questionnaire. Table 14 shows the results of this questionnaire. Additionally, we asked participants how helpful the CMAs were for the reflection activity. The students rated it as 51% for the heatmap, 70% for the gameplay recording, 57% for the eye movements superimposed on the gameplay screen, and 61% for the leading questions.

Table 14. Result of questionnaire about the reflective activity

Question	Option
If the reflection activity was helpful, how did it help?	· To plan and improve the next performance (100%) · To realize how I lack knowledge in the ABCDE method (94%) · To get familiarized with the game functions (94%) · To better understand the ABCDE method (64%) · Others (17%)
If the reflection activity was not helpful, why was it?	· I couldn't evaluate my performance due to lack of knowledge and feedback (83%) · I was not comfortable with the think-aloud (25%) · Others (4%)

Note. Participants were able to choose multiple options for each question.

4. Discussion

This study investigated the effects of the reflective pause on different aspects of performance and learning in simulation training. To make the reflective pause effective, we developed three types of CMAs: prompts, cues, and leading questions. We prompted three times of pauses between three part-task clusters and handover reporting in an extreme scenario for medical training. In the RP condition, participants could use cues and leading questions for their reflection. Using quantitative measures, we found that two aspects of performance improved through reflective pauses: Cognitive load is decreased, and the domain-general aspects of performance (vigilance on patient status) are enhanced. These effects appeared only in the later stage of the learning process, improving the performance in the third part-task cluster.

The first hypothesis (H1) assumed that the reflective pause decreases cognitive load in the intense scenario, as the participants could reinstate their working memory through reflection. This was supported by pupillometry, showing smaller pupil increase in the RP condition, while the subjective ratings did not show any significant difference between the conditions. We consider H1 to be at least partly confirmed since pupillometry shows higher sensitivity in measuring cognitive load than subjective ratings under two circumstances: when the task environment is highly dynamic (e.g., computer-based simulation), and when the task includes the management of emotions (e.g., controlling heighted stress in emergency situations).[8,29]

Interestingly, the effects of reflective pauses existed only in the third and last part-task cluster, signifying some latency in the effects. We postulate two explanations for this latency. First, an adaptation period is required to learn how to reflect using the CMAs.

Self-reflection is a metacognitive activity, a higher cognitive process that demands substantial resources in working memory resources.[69] Reportedly, these metacognitive activities can impose additional cognitive load.[8] On top of this, as participants were required to learn how to use the CMAs and process the novel information from them, it is likely that the participants were having trouble in the earlier stages to adapt themselves to the new tools. From the additional analysis on the baseline of each measure, we discovered significantly higher pupil increase at the first cluster in RP. We deem that this result indicates the participants' challenge against the adaptation, foreshadowing the latency.

The second explanation for the latency stems from the dynamic nature of tasks in critical fields. Task environments in these fields are continuously changing, where novel information gradually emerges as time progresses. The problems in these tasks are rarely caused by a single factor but are an accumulation of minor and latent errors.[70] Naturally, practitioners tend to first focus on task completion itself rather than reflecting on performance, while the signals of errors accumulate to a tipping point that composes an integral perception of a problem.[71,72] It is probable that information necessary for reflection was not fully accumulated for the participants at the first cluster, resulting in a lack of significant effects at that stage. Schmutz et al.[2] reported that reflection activity becomes more robust in the later stage of performance in extreme task environments, which is fully in line with our findings. This latency and the requirement of an adaptation period should be studied further in the future, for instance, by testing with a lengthened training duration where more time is provided to establish the effects on performance.

The second hypothesis (H2) expected the reflective pause to improve the domain-specific primary performance. Its theoretical assumption was that, during the reflective pause, participants could recap the current problems and develop creative solutions by interconnecting domain-specific knowledge in long-term memory (i.e., schemas). Contrary to our expectation, there was no significant difference in the primary performance between RP and No-RP in any cluster. We suppose that the students could succeed in calming down and recapping, yet could not arrive at the development of solutions to improve their future performance. They might have been struggling with identifying the gap between their problem states and goal states, not knowing how to improve their performance. The result of the questionnaire about the CMAs supports this explanation: When asked why the reflection activity was less helpful, 83% of students said they could not evaluate their performance due to lack of knowledge and feedback.

Feedback is the information that indicates whether points-of-improvement exist.[73] Provision of feedback appears to be crucial for reflection particularly for critical fields. In these fields, nonrecurrent aspects and uncertainty are prominent in tasks, thus the problem-solving processes are schema-based, requiring creative solutions.[74] Hence, feedback that can affect the schemas becomes essential to improve future performance.

Bardach et al.[25] reported that combining reflection and feedback is more effective than using only either reflection or feedback, which is in line with our explanation.

Although we intentionally excluded the feedback process from our study to identify the exact effects of reflection, future research should compare this setup to that with feedback. A necessary condition of feedback to foster reflection is to provide example solutions of experts or peers and ask learners to critically compare their own solutions with the example.[40] We suggest that such example solutions should be provided during reflection to make the reflection process fully effective. Additionally, more sensitive measure of domain-specific performance (e.g., diagnosis and intervention) can be explored further. Although game score is a widely used assessment of performance, it can still have limited sensitivity as it cannot account for the quality of primary performance in a global manner.[75,76]

The third hypothesis (H3) assumed that the reflective pause enhances domain-general secondary performance. As the optimization of cognitive load creates additional space for secondary load, participants could devote more attention to situational changes, enhancing the vigilance on patient status. In the RP condition, the transition rate across the vital signs monitor was significantly higher in the third cluster, supporting H3 and its assumptions. Transition rate has been known as an effective measure that can indicate the level of situational awareness and expertise in diverse medical domains.[57,77,78] Our findings demonstrate that it can also measure behavioral change in visual attention in simulation training. Besides, we consider that presenting participants' eye movements as a cue for reflection (i.e., heatmap and eye movements superimposed on the gameplay) helped to realize this change. Future research could investigate whether and how the use of visual cues can be an effective educational technique that facilitates reflection and behavioral changes in performance.

The last hypothesis (H4) assumed the improvement in schema structure thanks to the interconnection with long-term memory during the reflective pause. H4 was not corroborated, probably due to similar reasons as for H2. Due to the lack of feedback, participants could not properly diagnose their own schema nor identify whether points-of-improvement existed. Thus, despite the interconnection with long-term memory, learning could not be promoted, resulting in no improvement in schema structure. Future research could investigate how the combination of feedback and reflection affects learning as well as performance. Moreover, it should better be tested over a lengthened period of training time to find the longitudinal effects on learning, since the observed difference was in the expected direction.

The exploration of the reflection in RP yielded additional insights for future studies. The length of reflection was positively correlated with several performance measures (i.e., Paas scale, game score, and transition rate). We deem that this length merely represents the quantitative aspects of performance, such as the number of actions taken. Especially when using cues for reflection, description of action tends to become richer in comparison with when not using any cues.[48] If a participant carried out more actions during the simulation, their verbalization might become lengthier with their memory stimulated by cues, increasing perceived mental effort and game score in general. The transition rate's correlation with the previous cluster could mean that participants were simply verbalizing what they were looking at in the previous cluster, with scant connection with the performance in the next cluster. Future studies could investigate qualitative aspects of reflection by coding participants' verbalizations, to deepen the understanding of the reflection activity during pausing.

The results of our study clearly contribute to the existent debate on which method of in-action reflection (e.g., when and how to prompt reflection) is more effective for technology-enhanced learning.[20-22] There is no one good method that fits all different task environments: To make in-action reflection effective, proper CMAs should be provided through task analysis and specification of learning goals. For instance, in a computer-based simulation with an intense resuscitation scenario, reflection can be prompted between part-tasks clustered via task analysis, being provided with cues and leading questions for performance on patient safety. Additionally, this study demonstrates how physiological measures such as eye-tracking can be invaluable by revealing changes in cognitive processes in real time during performance. These contributions are in the line of Taub et al.'[21] review that called for definition of constructs in reflection-prompting methods and multimodal learner data in game-based learning.

This study opens new possibilities for research on in-action reflection in simulation-based learning. Using multiple CMAs and systematically categorizing performance measures, we have provided evidence for the benefits of the reflective pause. Its positive effects on managing heightened cognitive load and increasing vigilance on situational changes are clear advantages for professional skills in critical fields. Based on our findings, future research can be developed by investigating relevant topics, for instance, the effects of CMAs on learning, quality versus quantity of reflection, when and how to combine reflection and feedback, and how to develop an effective instructional design that takes the reflective pause into account.

From a practical view, designers of educational computer systems can immediately implement the reflective pause in their systems (e.g., serious games, computer-based simulation) by applying prompts, cues, and leading questions based on educational theories and task analysis. With these well-designed CMAs, their systems can effectively promote the complex skill of taking reflective pauses, creating extra learning opportunities just-in-time. Even without any feedback, the use of reflective pauses can already decrease learners' cognitive load in extreme scenarios, and encourage learners to be more circumspect on situational changes. As shown by the results of our questionnaire, students' perceived utility of the reflective pause and the CMAs is positive, which means the barriers should be low against the use of this technique in computer-based task environments. However, designers and educators should take precautions when implementing reflective pauses: familiarizing students with the reflective activity using CMAs takes time and effort, and their expected positive effects probably only show up in the later stage of the learning process. Moreover, the quality of reflection should also be considered, as its quantity can merely represent the number of actions taken and hardly affects the future behavior.

We recognize several limitations in our study. First, although we had built the theoretical background for the effects of the different types of CMAs, the effects of each type were not independently studied. It is possible that one type of aids (e.g., game replay, the add-on eye movements, heatmap) had positive effects while the others did not. Second, the quality, as opposed to quantity, of reflection and its correlation with performance and learning was not studied. Third, we used one specific scenario in medical domain, customizing CMAs for this scenario. To generalize our findings, development of other performance measures and CMAs specialized to different scenarios and domains should follow. Forth, since our experiment was conducted in a controlled lab environment, more factors from reality might have been excluded. For instance, time pressure during pauses can play a significant role in reality. Future research can study this time management aspect of the reflective pause.

To our knowledge, this study is the first attempt to identify tangible effects of the reflective pause in simulation-based training of complex skills in critical domains. Reflection during performance in extreme task environments can appear to be contradictory and uncomfortable for practitioners. When facing an emergency, they first jump into the task and even a brief moment of reflection can be seen as a 'luxury'. Therefore, researchers and educators should be even more pronounced in breaking a lance for the reflective pause. Being proven both theoretically and empirically, it enhances performance and the safety level. Moreover, the reflective pause can create learning opportunities to bolster lifelong learning for professionals, if implemented within a well-designed simulation training with enriched cognitive and metacognitive support.

References

1. Klein KJ, Ziegert JC, Knight AP, Xiao Y. Dynamic Delegation: Shared, Hierarchical, and Deindividualized Leadership in Extreme Action Teams. *Adm Sci Q.* 2006;51(4):590-621. 10.2189/asqu.51.4.590
2. Schmutz JB, Lei Z, Eppich WJ, Manser T. Reflection in the heat of the moment: The role of in-action team reflexivity in health care emergency teams. *Journal of Organizational Behavior.* 2018;39(6): 749-765. 10.1002/job.2299
3. Ishak AW, Ballard DI. Time to Re-Group: A Typology and Nested Phase Model for Action Teams. *Small Group Research.* 2012;43(1):3-29. 10.1177/1046496411425250
4. Clapper TC, Leighton K. Incorporating the reflective pause in simulation: A practical guide. *The Journal of Continuing Education in Nursing.* 2020;51(1):32-38. 10.3928/00220124-20191217-07
5. St-Martin L, Patel P, Gallinger J, Moulton CA. Teaching the slowing-down moments of operative judgment. *Surgical Clinics.* 2012;92(1):125-135. 10.1016/j.suc.2011.12.001
6. Moulton CA, Epstein RM. Self-monitoring in surgical practice: Slowing down when you should. In: Fry H. KR, ed. *Surgical Education*. Springer, Dordrecht; 2011:169-182.
7. Lee JY, Szulewski A, Young JQ, Donkers J, Jarodzka H, Van Merriënboer JJG. The Medical Pause: Importance, Processes, and Training. *Med Educ.* 2021. 10.1111/medu.14529
8. Lee JY, Donkers J, Jarodzka H, Sellenraad G, Van Merriënboer JJG. Different effects of pausing on cognitive load in a medical simulation game. *Comput Human Behav.* 2020;110:106385. 10.1016/j.chb.2020.106385
9. Hasler BS, Kersten B, Sweller J. Learner control, cognitive load and instructional animation. *Appl Cogn Psychol.* 2007;21(6):713-729. 10.1002/acp.1345
10. Schmutz JB, Kolbe M, Eppich WJ. Twelve tips for integrating team reflexivity into your simulation-based team training. *Med Teach.* 2018;40(7):721-727. 10.1080/0142159x.2018.1464135
11. Trowbridge RL. Twelve tips for teaching avoidance of diagnostic errors. *Med Teach.* 2008;30(5): 496-500. 10.1080/01421590801965137
12. Kelly JJ, Farley H, O'Cain C, et al. A survey of the use of time-out protocols in emergency medicine. *The Joint Commission Journal on Quality and Patient Safety.* 2011;37(6):285-288. 10.1016/S1553-7250(11)37036-5
13. Rydenfalt C, Johansson G, Odenrick P, Akerman K, Larsson PA. Compliance with the WHO Surgical Safety Checklist: deviations and possible improvements. *Int J Qual Health Care.* 2013; 25(2):182-187. 10.1093/intqhc/mzt004
14. Huang T-T. *The effects of types of reflective scaffolding and language proficiency on the acquisition of physics knowledge in a game-based learning environment*, New York University; 2016.
15. Nelson BC. Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *Journal of Science Education and Technology.* 2007;16(1):83-97.
16. Ke F. Computer games application within alternative classroom goal structures: cognitive, metacognitive, and affective evaluation. *Educational Technology Research and Development.* 2008; 56(5-6):539-556. 10.1007/s11423-008-9086-5

17. Viberg O, Khalil M, Baars M. Self-regulated learning and learning analytics in online learning environments: A review of empirical research. Paper presented at: Proceedings of the tenth international conference on learning analytics & knowledge2020.

18. Fiorella L, Mayer RE. Paper-based aids for learning with a computer-based game. *J Educ Psychol.* 2012;104(4):1074.

19. Ter Vrugte J, De Jong T, Wouters P, Vandercruysse S, Elen J, Van Oostendorp H. When a game supports prevocational math education but integrated reflection does not. *J Comput Assisted Learn.* 2015;31(5):462-480. 10.1111/jcal.12104

20. Kori K, Pedaste M, Leijen Ä, Mäeots M. Supporting reflection in technology-enhanced learning. *Educational Research Review.* 2014;11:45-55. 10.1016/j.edurev.2013.11.003

21. Taub M, Azevedo R, Bradbury AE, Mudrick NV. Self-regulation and reflection during game-based learning. *Handbook of Game-Based Learning.* 2020:239-262.

22. Yang X, Liu Y. Supporting Students' Reflection in Game-Based Science Learning: A Literature Review. In: Li R. CSKS, Iwasaki C., Kwok LF., Kageto M. , ed. *Lecture Notes in Computer Science.* Springer International Publishing; 2021:119-131.

23. Azevedo R, Mudrick NV, Taub M, Bradbury AE. Self-regulation in computer-assisted learning systems. 2019.

24. Azevedo R, Taub M, Mudrick NV, Martin S, Grafsgaard J. Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. *Handbook of self-regulation of learning and performance.* 2018;2.

25. Bardach L, Klassen RM, Durksen TL, Rushby JV, Bostwick KC, Sheridan L. The Power of Feedback and Reflection: Testing an Online Scenario-based Learning Intervention for Student Teachers. *Computers & Education.* 2021:104194. 10.1016/j.compedu.2021.104194

26. Jansen RS, van Leeuwen A, Janssen J, Conijn R, Kester L. Supporting learners' self-regulated learning in Massive Open Online Courses. *Computers & Education.* 2020;146:103771. https://doi.org/10.1016/j.compedu.2019.103771

27. Sweller J, Van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev.* 2019;31(2):261-292. 10.1007/s10648-019-09465-5

28. Fraser KL, Ayres P, Sweller J. Cognitive Load Theory for the Design of Medical Simulations. *Simulation in Healthcare.* 2015;10(5):295-307. 10.1097/SIH.0000000000000097

29. Naismith LM, Cavalcanti RB. Validity of cognitive load measures in simulation-based training: a systematic review. *Acad Med.* 2015;90(11):S24-S35. 10.1097/ACM.0000000000000893

30. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. *Med Teach.* 2009; 31(8):685-695. 10.1080/01421590903050374

31. Wiegmann D, Faaborg T, Boquet A, Detwiler C, Holcomb K, Shappell S. *Human error and general aviation accidents: A comprehensive, fine-grained analysis using HFACS.* FEDERAL AVIATION ADMINISTRATION OKLAHOMA CITY OK CIVIL AEROMEDICAL INST;2005.

32. Institute of Medicine. *To Err is Human: Building a Safer Health System.* Vol 6. Washington, DC: National Academies Press; 2000.

33. Moulton CA, Regehr G, Lingard L, Merritt C, MacRae H. Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. *Acad Med.* 2010;85(10):1571-1577. 10.1097/ACM.0b013e3181f073dd

34. Bloom BS. *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill Book Company; 1971.
35. Heitz RP. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front Neurosci.* 2014;8(150):1-19. 10.3389/fnins.2014.00150
36. Moulton CA, Regehr G, Mylopoulos M, Macrae HM. Slowing Down When You Should: A New Model of Expert Judgment. *Acad Med.* 2007; 82(Suppl): S109-S116. 10.1097/acm.0b013e3181405a76
37. Nelson TO. Metamemory: A theoretical framework and new findings. *Psychology of learning and motivation.* 1990;26:125-173. 10.1016/S0079-7421(08)60053-5
38. De Bruin ABH, Dunlosky J, Cavalcanti RB. Monitoring and regulation of learning in medical education: the need for predictive cues. *Med Educ.* 2017;51(6):575-584. 10.1111/medu.13267
39. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001
40. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
41. Butler DL, Winne PH. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research.* 1995;65(3):245-281. 10.3102/00346543065003245
42. Chiu JL, Chi MT. Supporting self-explanation in the classroom. In: Benassi VA, Overson CE, Hakala CM, eds. *Applying science of learning in education: Infusing psychological science into the curriculum.*2014:91-103.
43. Spanjers IAE, Van Gog T, Wouters P, Van Merrienboer JJG. Explaining the segmentation effect in learning from animations: The role of pausing and temporal cueing. *Computers & Education.* 2012;59(2):274-280. 10.1016/j.compedu.2011.12.024
44. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. *Educ Psychol.* 2003;38(1):43-52. 10.1207/S15326985ep3801_6
45. Russo JE, Johnson EJ, Stephens DL. The validity of verbal protocols. *M&C.* 1989;17(6):759-769. 10.3758/bf03202637
46. Kuusela H, Paul P. A Comparison of Concurrent and Retrospective Verbal Protocol Analysis. *The American Journal of Psychology.* 2000;113(3):387. 10.2307/1423365
47. Van Someren M, Barnard Y, Sandberg J. *The think aloud method: a practical approach to modelling cognitive processes.* Londen: Academic Press; 1994.
48. Van Gog T, Paas F, Van Merrienboer JJ, Witte P. Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *J Exp Psychol Appl.* 2005; 11(4):237-244. 10.1037/1076-898X.11.4.237
49. Lauwereyns J, D'Ydewalle G. Knowledge acquisition in poetry criticism: the expert's eye movements as an information tool. *Int J Hum Comput Stud.* 1996;45(1):1-18. 10.1006/ijhc.1996.0039
50. Susac A, Bubic A, Kaponja J, Planinic M, Palmovic M. Eye movements reveal students' strategies in simple equation solving. *International Journal of Science and Mathematics Education.* 2014; 12(3):555-577. 10.1007/s10763-014-9514-4
51. Hansen JP. The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychol (Amst).* 1991;76(1):31-49. 10.1016/0001-6918(91)90052-2

52. Russo JE. A software system for the collection of retrospective protocols prompted by eye fixations. *Behavior Research Methods & Instrumentation.* 1979;11(2):177-179. 10.3758/bf03205643
53. Sweller J, Kirschner PA, Clark RE. Why Minimally Guided Teaching Techniques Do Not Work: A Reply to Commentaries. *Educ Psychol.* 2007;42(2):115-121. 10.1080/00461520701263426
54. Kirschner F, Paas F, Kirschner PA. Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Comput Human Behav.* 2009;25(2):306-314. 10.1016/j.chb.2008.12.008
55. Collins A, Ferguson W. Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educ Psychol.* 1993;28(1):25-42. 10.1207/s15326985ep2801_3
56. Renkl A. Worked-out examples: instructional explanations support learning by self-explanations. *Learning and Instruction.* 2002;12(5):529-556. 10.1016/s0959-4752(01)00030-5
57. Lee JY, Donkers J, Jarodzka H, van Merriënboer JJ. How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Comput Human Behav.* 2019;99:268-277. 10.1016/j.chb.2019.05.035
58. *AbcdeSIM* [computer program]. Rotterdam: VirtualMedSchool; 2012.
59. Holmqvist K, Andersson R. *Eye-tracking: A comprehensive guide to methods, paradigms and measures.* Lund, Sweden: Lund Eye-Tracking Research Institute; 2017.
60. Paas FGWC. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics - a Cognitive-Load Approach. *J Educ Psychol.* 1992;84(4):429-434. 10.1037/0022-0663.84.4.429
61. Patton MQ. Enhancing the quality and credibility of qualitative analysis. *Health Serv Res.* 1999; 34:1189-1208. https://pubmed.ncbi.nlm.nih.gov/10591279.
62. Schmidt HG, Rikers RMJP. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Med Educ.* 2007:1133-1139. 10.1111/j.1365-2923.2007.02915.x
63. Haig KM, Sutton S, Whittington J. SBAR: A Shared Mental Model for Improving Communication Between Clinicians. *The Joint Commission Journal on Quality and Patient Safety.* 2006;32(3): 167-175. 10.1016/s1553-7250(06)32022-3
64. *R: A language and environment for statistical computing* [computer program]. R Foundation for Statistical Computing; 2019.
65. Szulewski A, Gegenfurtner A, Howes DW, Sivilotti ML, van Merriënboer JJ. Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education.* 2017;22(4):951-968. 10.1007/s10459-016-9725-2
66. Van Rossum G, Drake FL. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace; 2009.
67. Tukey JW. *Exploratory data analysis.* Vol 2. Reading, Mass: Addison-Wesley; 1977.
68. Bates D, Machler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software.* 2015;67(1):1-48. 10.18637/jss.v067.i01.
69. Seufert T. The interplay between self-regulation in learning and cognitive load. *Educational Research Review.* 2018;24:116-129. 10.1016/j.edurev.2018.03.004
70. Lei Z, Waller MJ, Hagen J, Kaplan S. Team Adaptiveness in Dynamic Contexts. *Group & Organization Management.* 2016;41(4):491-525. 10.1177/1059601115615246
71. Rudolph JW, Repenning NP. Disaster Dynamics: Understanding the Role of Quantity in Organizational Collapse. *Adm Sci Q.* 2002;47(1):1. 10.2307/3094889

72. Waller MJ, Uitdewilligen S. Talking to the room: Collective sensemaking during crisis situations. In: Roe RA, Waller MJ, Clegg SR, eds. *Time in organizational research*. Routledge; 2008:208-225.

73. Hattie J, Timperley H. The Power of Feedback. *Review of Educational Research*. 2007;*77*(1): 81-112. 10.3102/003465430298487

74. Van Merriënboer JJ. Perspectives on problem solving and instruction. *Computers & Education*. 2013;64:153-160. 10.1016/j.compedu.2012.11.025

75. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education*. 1996;1(3):227-233. 10.1007/BF00162920

76. Dankbaar ME, Stegers-Jager KM, Baarveld F, et al. Assessing the assessment in emergency care training. *PLoS One*. 2014;9(12). 10.1371/journal.pone.0114663

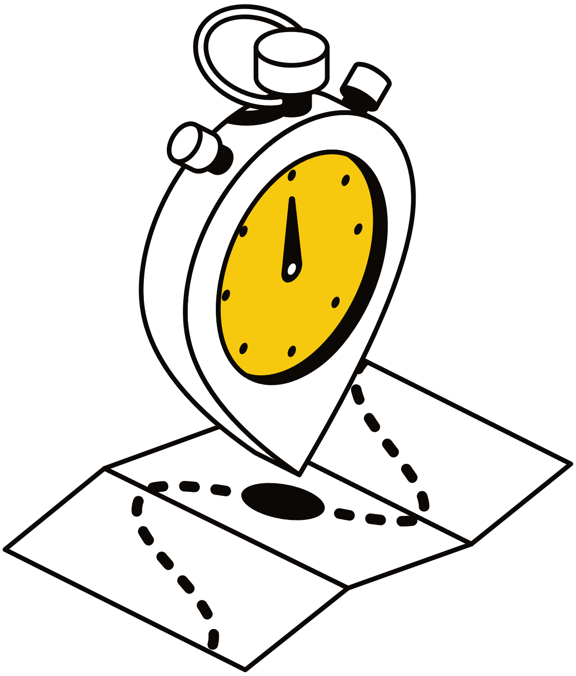
77. Aldekhyl S, Cavalcanti RB, Naismith LM. Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on Medical Education*. 2018;7(1):23-32. 10.1007/s40037-017-0392-7

78. Tien G, Atkins MS, Zheng B, Swindells C. Measuring situation awareness of surgeons in laparoscopic training. Paper presented at: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications; March, 2010.

Chapter 6

Measuring Cognitive Load in Virtual Reality Training via Pupillometry

This chapter is under revision as:
Lee JY, De Jong N, Donkers J, Jarodzka H, Van Merriënboer JJG. Measuring cognitive load in Virtual Reality training via pupillometry.



Abstract

Pupillometry is known as a reliable technique to measure cognitive load in learning and performance. However, its applicability to virtual reality (VR) environments, an emerging technology for simulation-based training, has not been validated. Specifically, the VR display causes light reflexes that confound task-evoked pupillary responses (TEPRs), impairing cognitive load measures. Through this pilot study, we validated whether task difficulty can predict cognitive load as measured by TEPRs corrected for the light reflex and if these TEPRs correlate with cognitive load self-ratings and performance. 14 students in health sciences performed observation tasks in two conditions: difficult versus easy tasks, whilst watching a VR scenario in home health care. Then, a cognitive load self-rating ensued. We used a VR system with a built-in eye-tracker and a photosensor installed to assess pupil diameter and light intensity during the scenario. Employing a method from the human-computer interaction field, we determined TEPRs by modeling the pupil light reflexes using a baseline. As predicted, the difficult task caused significantly larger TEPRs than the easy task. Only in the difficult task condition did TEPRs positively correlate with the performance measures. These results suggest that TEPRs are valid measures of cognitive load in VR training when corrected for the light reflex. It opens up possibilities to use real-time cognitive load for assessment and instructional design for VR training. Future studies should test our findings with a larger sample size, in various domains, involving complex VR functions such as haptic interaction.

Keywords: Virtual reality; medical training; simulation; cognitive load; physiological measure

1. Introduction

As the COVID-19 pandemic is rapidly changing the education experience, Virtual Reality (VR) has become a powerful alternative to physical simulation-based training.[1] VR creates immersive task environments that provoke a sense of presence, emotions, and engagement,[2,3] which allows for a favorable training environment with high fidelity of specific tasks.[4-6] Some studies have reported positive effects of VR training on learning outcomes and perceived usability.[7,8] However, key challenges still remain: instructional design or course structure for VR training have not been fully established,[6,9,10] and few studies have actually measured students' performance during VR training.[7] In order to build an effective instructional design for a new task environment, performance assessment and progress monitoring are fundamental.[11] This necessitates the development of real-time measures of learning including task performance and cognitive load.

1.1. Cognitive Load Theory

Cognitive load theory (CLT) provides a useful framework to analyze learning processes in simulation training.[12-15] CLT posits an "element interactivity" where heterogeneous processes in cognitive, affective, and social domains coincide in working memory,[13,16] which is particularly relevant to simulation training that requires multiple tasks to be performed simultaneously within complex environments,[12,17,18] Cognitive load refers to the imposition of these processes caused by given tasks, to working memory that has limited capacity.[13,19]

There are three types of cognitive load at a conceptual level: intrinsic cognitive load which reflects the complexity of a task and a learner's competency for performing the task, germane cognitive load that pertains to learning, and extraneous cognitive load that stems from distraction or suboptimal instructional design.[20] Among these three, intrinsic load can be measured experimentally by varying the level of task complexity.[21] If the sources of germane and extraneous load are kept constant (e.g., excluding the requirement for learning and causes of distraction), intrinsic cognitive load can be assumed to be equivalent to total cognitive load.[22] As intrinsic load can reflect learners' competency for performing a task, the level of cognitive load could correlate with performance outcomes. Studies have shown that this correlation can be either positive or negative, depending on characteristics of measurements or research context.[12,23]

If measured properly, cognitive load can be an effective indicator of performance, learning, and expertise, which in turn informs instructional design.[13,24] Thus, measuring cognitive load with a valid and reliable method has been an important issue in research on learning and instruction.[13,25] In general, three methods have been used to measure cognitive load: self-rating, secondary tasks, and psychophysiological indices.[12,25] In simulation training, psychophysiological indices are shown to be the most sensitive measure of the three.[12] Specifically, eye-tracking indices have been used as a cognitive load measure for decades,[26-28] demonstrating higher validity among other psychophysiological measures such as heart rate or heart rate variability.[12]

1.2. Using Pupillometry in Virtual Reality

Pupil dilation is a well-validated measure of cognitive load in simulation training.[24] A large body of literature has confirmed that pupil dilation correlates with cognitive demands imposed by tasks such as solving arithmetic problems or spelling difficult words.[29-33] At a physiological level, pupil dilation is known to be an involuntary response that reflects noradrenergic activity in the locus coeruleus which regulates arousal, mental activity, and emotion.[34,35] Since pupil dilation reflects emotional activity, it can be an effective measure of cognitive load caused by task difficulty especially when the task includes the management of emotion.[23] Moreover, pupil dilation may capture real-time changes in cognitive load more robustly than self-rating in dynamic environments such as computer-based simulation.[23]

Most modern eye-tracking systems can measure pupil diameter economically and non-invasively.[36] However, when using pupillometry to measure cognitive load in dynamic task environments, researchers should be wary of a major confounder, the pupillary light reflex. The cognitive effects on pupil size are notably small compared with the effects of light intensity. The light condition of experimental rooms and the stimuli can affect pupil size by 50%, while the changes in pupil size caused by cognitive processing for the given task, or task-evoked pupillary responses (TEPR), account for only about 5% [36]. In less dynamic task environments with static visual stimuli, researchers can control for light reflexes by maintaining a constant light level in the room and using a baseline image of the stimuli. However, these methods cannot be readily applied to VR simulation where light conditions in the visual field change dynamically and participants move around between different locations in the 3D world.

In fact, research that uses pupillometry to study cognitive processes during VR training is still in its early stage. Several challenges to this research can be recognized: (1) difficulty of access to VR systems connected with eye trackers, (2) lack of learning contents in VR training where task complexity can be scaled, and (3) technical challenges in determining TEPRs by correcting for the light reflex.[37] In the field of human-computer interaction, a few studies have shown that some cognitive processes such as emotion can be measured in VR through psychophysiological indices.[37,38]

This pilot study explores whether we could address the aforementioned challenges through a structured experimental setup. In this setup, we build a VR system with an eye tracker, develop a VR training environment where task complexity can be adjusted, and apply the technical methods to identify TEPRs. Three goals are specified for this study. Firstly, we determine TEPRs by employing a light reflex compensation technique developed by Chen et al..[37] Secondly, we validate these TEPRs as a cognitive load measure in VR training using two approaches: predictive validity that examines whether task difficulty as a factor of cognitive load predicts changes in TEPRs, and concurrent validity to test whether TEPRs positively correlate with other cognitive load measures such as self-rating. Thirdly, assuming the validity, we explore correlations between TEPRs and performance.

We establish two hypotheses in order to examine the validity of the measure and the correlations between cognitive load and performance through an experimental setup, which will be reflected in the following sections.

H1. In a VR training environment, difficult tasks increase cognitive load more than easy tasks.

H2. In a VR training environment, the level of cognitive load correlates with performance level.

2. Methods

2.1. Participants and Design

Fourteen undergraduate students (12 females; mean age 20.5; *SD* = 1.5) in Health Sciences were recruited at Maastricht University, the Netherlands. We used a within-subjects design with task difficulty as the single factor. Two different conditions were presented with the presentation order counterbalanced: easy task condition (ET) where participants performed a simple observation task whilst watching a VR scenario, and difficult task condition (DT) with high task complexity.

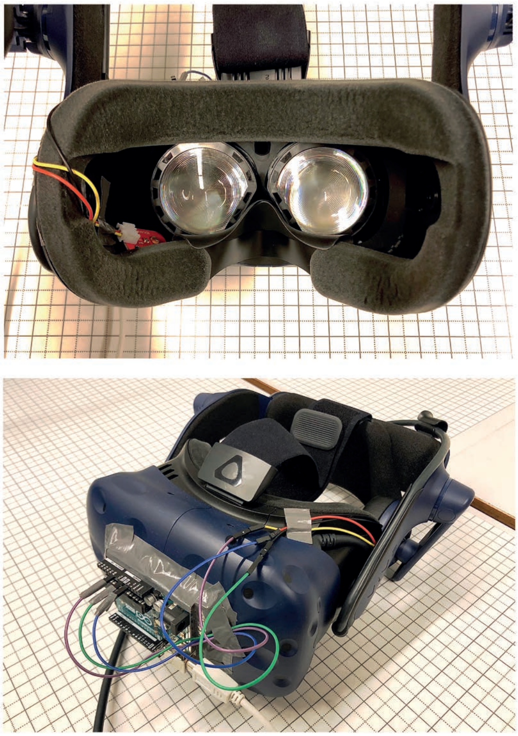
2.2. Materials and Technical Setup

A VR scenario for home health care was developed by a research team at Maastricht University. In this scenario, a healthcare provider visits a patient to deliver medical care and share social interaction. This scenario takes 9 minutes and is formatted for a 360-degree head-mounted display.

Two researchers active in health sciences designed the observation tasks and the assessment. In ET, a simple instruction was given: “Observe and report: What is the homecare provider doing in the scenario?” In DT, more detailed observations were requested: “What are the patient’s symptoms? Describe at least three symptoms; What are the strengths and weaknesses of the provider’s performance? Describe at least three for each strengths and weaknesses.” The observation results were reported in writing. We defined two performance measures: task score to represent the quality of the report,

and report length for the level of detail measured by word count. The two researchers scored the report, based on "Professional profile of registered nurses", a health sciences document commonly used in the Netherlands.
A personal computer (Intel Core i9-9900K, 32 GB RAM) ran the scenario, displaying it through an HTC Vive Pro Eye headset (2880 x 1600 pixels, 110-degree visual field). This headset has a built-in eye tracker with 120 Hz sampling rate that uses HTC SRanipal SDK as an interface (version 1.3.1.0, www.vive.com). Following Chen et al.'s method,[37] a photosensor (LDR sensor Iduino ST1107) was installed inside the headset to measure the light intensity of the display (the upper side of Figure 19). An Arduino board (Arduino UNO Rev3) connected the sensor to the computer via USB (the lower side of Figure 19). The WorldViz Vizard software (version 6.0, www.worldviz.com) was used to arrange stimuli presentation and data recording. We used the VR Eye-tracking Analytics Lab package to synchronize the data from both the eye-tracker and the photosensor.

Figure 19. Photosensor installed inside the VR headset



Note. On the upper side, a photosensor was installed inside the headset to measure the light intensity of the display. On the lower side, the Arduino board was mounted on the outside of the headset to connect the sensor to the computer via USB.

To model the light reflexes of individuals, we used a baseline stimulus, a VR scene of an empty room. The light intensity of this scene was arranged to increase in 10 stages where each stage takes 5 seconds. For the self-rating of cognitive load, we employed the widely used 9-point Paas Scale with the value 1 representing the lowest mental effort and the value 9 representing the highest.[39]

2.3. Procedure

Individual sessions consisted of two trials, ET and DT in counterbalanced order. Initially, participants were presented with instructions about the task. Next, they were positioned to stand at the center of the VR area, then put on the VR headset. After a 5-point calibration and the baseline session, the scenario was presented. Participants could rotate their heads or walk around in the VR area during the scenario. After the scenario, they reported their task results, and self-rated their cognitive load for the task on the Paas Scale.

2.4. Data Analysis

The raw data included timestamps, light intensity, and pupil diameter. Using the baseline data, we identified the dependence between light intensity and pupil diameter for each trial of the participants. Following Watson and Yellott's work that presented a formula describing light-adapted pupil size, a third-degree polynomial was fitted to the pupil diameter versus light intensity.[40] This resulted in a trend line formula for each trial. Applying this formula, pupil dilation caused by light reflexes was predicted and subtracted from the raw pupil diameter. We assumed the rest of the pupil diameter to be TEPRs. These TEPRs were averaged over time for each trial, then the difference between ET and DT was tested via a dependent sample *t* test.

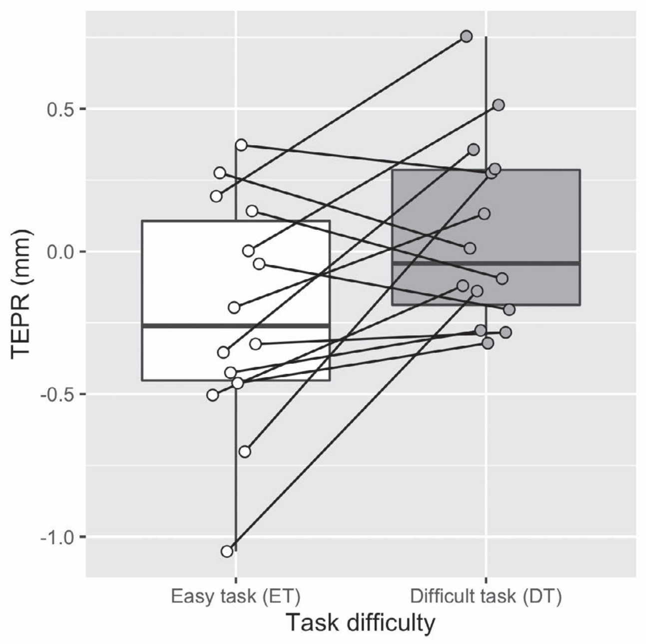
For calculating correlations, the two conditions of ET and DT were averaged for each participant, except when focusing on each condition. We calculated the Pearson correlation between continuous normal datasets (i.e., TEPRs and the report length) while using Kendall's method for the rest of the datasets. The task scores were averaged between the two raters. The inter-rater reliability was examined ($r_T = 0.80$). R (version 3.5.1, R Core Team, 2019) was used for statistical analysis. We considered $p < .05$ to be statistically significant.

3. Results

3.1. TEPRs as a Cognitive Load Measure

Figure 20 demonstrates the change in TEPRs between ET and DT for all 14 participants. As predicted, TEPRs were significantly larger in DT (mean = 0.06, SD = 0.33) than in ET (mean = -0.22, SD = 0.40) (p = .01). The model fits of the third-degree polynomial were sufficient to show that the method can function properly (R^2 mean = 0.86, SD = 0.09).

Figure 20. The changes in TEPRs between ET and DT for all participants



Note. The boxplots depict the quartile-based distribution of individual samples in each condition.

The Paas Scale significantly correlated with TEPRs (r_T = 0.51, p = .01). The self-rating score was higher in DT (median = 5, range = 2-7) than in ET (median = 4, range = 2-6), yet the difference between the conditions was not statistically significant (p = .07).

3.2. Correlations between TEPRs and Performance

When averaging ET and DT, TEPRs did not significantly correlate with performance. However, when focusing on either condition, DT showed significant correlations. In DT, TEPRs positively correlated with task score (r_T = 0.45, p = .03) and report length (r = 0.58, p = .03), while the Paas Scale did with task score only (r_T = 0.56, p = .01) (Table 15).

4. Discussion

The present study tested the validity of TEPRs as a measure of cognitive load in a VR training environment. We have developed a VR learning environment where task complexity can be adjusted, established a VR system with an eye tracker, and controlled for the light reflex to determine TEPRs. We have confirmed the predictive validity as cognitive load increased with task difficulty, and the concurrent validity as the cognitive load correlated with the self-ratings. Additionally, performance measures correlated with TEPRs only in DT.

Table 15. Correlations between TEPRs, Paas Scale, and the two performance measures in DT only

Variable	TEPR	Paas Scale	Task score
Paas Scale	0.54 (.01)		
	[.20, .89]		
Task score	0.45 (.03)	0.56 (.01)	
	[.13, .77]	[.27, .84]	
Report length	0.58 (.03)	0.09 (.69)	0.37 (.07)
	[.07, .85]	[-.21, .38]	[.02, .72]

Note. p values are presented in parentheses. Values in square brackets indicate the 95% confidence interval for each correlation.

The first hypothesis (H1) assumed that cognitive load is higher in difficult tasks than in easy tasks in a VR training environment. We found a significant impact of task complexity on cognitive load by using TEPRs, whereas the cognitive load self-rating did not show statistically significant effects. An explanation for the higher sensitivity of TEPRs compared with the self-rating might reside in the unique characteristics of VR task environments that provoke dynamic emotions through immersion.[37] While self-rating scales have been developed for classroom-based settings and depend solely on participants' judgment, pupil dilation is an involuntary response that reflects arousal and emotions.[36] Simulation training often involves affective and social tasks that require the management of emotions. Our observation tasks also included detecting the healthcare provider's empathy for the patient. In such task environments, pupillometry could be a more sensitive measurement of cognitive load than self-rating. This finding is consistent with previous research in simulation-based training.[12,23] We suggest that more future research is needed to study this effect of emotional engagement on cognitive load in VR training.

Consequently, we largely confirm H1 based on the higher sensitivity of TEPRs as a cognitive load measure in VR training environments. This supports TEPR's predictive validity as a cognitive load measure. Although the cognitive load self-rating showed lower sensitivity, it significantly correlated with TEPRs, which endorses TEPR's concurrent validity as a cognitive load measure.

The second hypothesis (H2) posited that the level of cognitive load correlates with performance level in a VR training environment. Using TEPRs, we found that cognitive load positively correlates with the two performance measures (i.e., task score and report length), but only in DT condition. Here we suspect there is a ceiling effect in ET that might have caused a weak discrimination in performance measures. We recommend that future studies should include a task analysis beforehand in order to prevent scale attenuation effects. Using the cognitive load self-rating, we found significant correlations only in DT as well, but only with one performance measure (i.e., task score). We consider that this might be related to the lower sensitivity of self-rating, which was explained through the results of H1.

Concerning the positive direction of the correlation, it is likely that the sources of cognitive load in this study were rather beneficial to performance. For instance, they might have been stimulating information processing rather than harming performance. It appears that the task environment in this study is not as overwhelming as other simulation environments. The scenario was slow-paced, and the observation tasks were passive without any psychomotor skills required. Studies have shown that correlations between cognitive load and performance varied from positive to negative across different research contexts.[12] While this inconsistency may partly stem from measurement limitations,[12] we argue that the nature of the factors that caused cognitive load determines the direction of the correlations. If the factors are negative to performance (e.g., distraction) or make the total cognitive load exceed working memory capacity, cognitive load should be inversely proportional to performance. If the factors are positive (e.g., so-called *germane* cognitive load,[13,21] self-regulation[23,41]), cognitive load can positively correlate with performance. Again, we emphasize the importance of preceding task analysis to define the sources of cognitive load in future studies.

The validity and utility of TEPRs found by the present study open up new possibilities to improve research on VR training and instructional design for VR environments. Researchers may expand this finding to more diverse VR training environments, investigate the potential of pupillometry to assess cognitive processes during VR training, and test if this new measure can be used to evaluate performers' expertise. For educators who search for an effective assessment tool for VR training, TEPRs can be a good option that provides an objective indicator of performers' competence to manage situational and emotional challenges in complex environments. This assessment might compensate for the lacking discriminatory power of traditional performance measures (e.g., questionnaires, checklists), improving instructional design and training programs for VR environments.

Our study has several limitations. First, as a pilot study, we used a small sample of participants. This might reduce the generalizability of our findings. Second, the scenario included only one domain, i.e., home health care. Our findings should be tested if they are applicable to more diverse domains. Third, the VR content we used was formatted for a 360-degree head-mounted display, which is only one type of VR technology. Future studies should examine our methods in more advanced settings such as 3D-rendered VR with haptic interaction. For these studies, a more careful control for confounding factors should be required as various sensory modalities are involved in such complex environments. Lastly, other confounding factors in pupillometry such as pupil foreshortening error (i.e., the influence of gaze position on pupil size)[36,42] were not corrected, due to a lack of corresponding methods for VR environments.

To our knowledge, this study is the first to show the validity of TEPRs as a cognitive load measure in VR healthcare training. The hidden potential of using VR training lies in the utility of datasets from diverse sources such as eye-tracking, which provides rich information about training development. Continued study is needed to improve the understanding of these datasets and make VR healthcare training more effective.

References

1. Woolliscroft JO. Innovation in Response to the COVID-19 Pandemic Crisis. *Acad Med.* 2020; 95(8):1140-1142. 10.1097/acm.0000000000003402
2. Slater M. Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence.* 1999;8(5):560-565.
3. Riva G, Mantovani F, Capideville CS, et al. Affective Interactions Using Virtual Reality: The Link between Presence and Emotions. *CyberPsychol Behav.* 2007;10(1):45-56. 10.1089/cpb.2006.9993
4. Pellas N, Kazanidis I, Palaigeorgiou G. A systematic literature review of mixed reality environments in K-12 education. *Education and Information Technologies.* 2019:1-40.
5. Hite R, Jones M, Childers G, et al. Investigating potential relationships between adolescents' cognitive development and perceptions of presence in 3-D, haptic-enabled, virtual reality science instruction. *Journal of Science Education and Technology.* 2019;28(3):265-284.
6. Merchant Z, Goetz ET, Cifuentes L, Keeney-Kennicutt W, Davis TJ. Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education.* 2014;70:29-40. 10.1016/j.compedu.2013.07.033
7. Pellas N, Dengel A, Christopoulos A. A Scoping Review of Immersive Virtual Reality in STEM Education. *IEEE Transactions on Learning Technologies.* 2020;13(4):748-761.
8. Pantelidis P, Chorti A, Papagiouvanni I, et al. Virtual and Augmented Reality in Medical Education. In: Tsoulfas G, ed. *Medical and Surgical Education: Past, Present and Future.* London, UK: IntechOpen; 2018:77-97.
9. Jensen L, Konradsen F. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies.* 2018;23(4):1515-1529.
10. Fertleman C, Aubugeau-Williams P, Sher C, et al. A Discussion of Virtual Reality As a New Tool for Training Healthcare Professionals. *Frontiers in Public Health.* 2018;6. 10.3389/fpubh.2018.00044
11. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
12. Naismith LM, Cavalcanti RB. Validity of cognitive load measures in simulation-based training: a systematic review. *Acad Med.* 2015;90(11):S24-S35. 10.1097/ACM.0000000000000893
13. Sweller J, Van Merriënboer JJ, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev.* 2019;31(2):261-292. 10.1007/s10648-019-09465-5
14. Fraser KL, Ayres P, Sweller J. Cognitive Load Theory for the Design of Medical Simulations. *Simulation in Healthcare.* 2015;10(5):295-307. 10.1097/SIH.0000000000000097
15. Moreno RE, Park B. Cognitive load theory: Historical development and relation to other theories. 2010.
16. Young JQ, O'Sullivan PS, Ruddick V, Irby DM, Cate OT. Improving Handoffs Curricula: Instructional Techniques from Cognitive Load Theory. *Acad Med.* 2017;92(5):719. 10.1097/acm.0000000000001664

17. Haji FA, Rojas D, Childs R, De Ribaupierre S, Dubrowski A. Measuring cognitive load: performance, mental effort and simulation task complexity. *Med Educ.* 2015;49(8):815-827. 10.1111/medu.12773
18. Young JQ, Van Dijk SM, O'Sullivan PS, Custers EJ, Irby DM, Ten Cate O. Influence of learner knowledge and case complexity on handover accuracy and cognitive load: results from a simulation study. *Med Educ.* 2016;50(9):969-978. 10.1111/medu.13107
19. Van Merrienboer JJ, Sweller J. Cognitive load theory and complex learning: Recent developments and future directions. *Educ Psychol Rev.* 2005;17(2):147-177.
20. Sweller J. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ Psychol Rev.* 2010;22(2):123-138.
21. Szulewski A, Howes D, Van Merriënboer JJG, Sweller J. From Theory to Practice: The Application of Cognitive Load Theory to the Practice of Medicine. *Acad Med.* 2020;Publish Ahead of Print. 10.1097/acm.0000000000003524
22. Ayres P. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction.* 2006;16(5):389-400. 10.1016/j.learninstruc.2006.09.001
23. Lee JY, Donkers J, Jarodzka H, Sellenraad G, Van Merriënboer JJG. Different effects of pausing on cognitive load in a medical simulation game. *Comput Human Behav.* 2020;110:106385. 10.1016/j.chb.2020.106385
24. Szulewski A, Roth N, Howes D. 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. *Acad Med.* 2015;90(7):981-987. 10.1097/Acm.0000000000000677
25. Korbach A, Brünken R, Park B. Differentiating different types of cognitive load: A comparison of different measures. *Educ Psychol Rev.* 2018;30(2):1-27. 10.1007/s10648-017-9404-8
26. Glaholt MG. Eye tracking in the cockpit: a review of the relationships between eye movements and the aviators cognitive state. 2014.
27. Rosch JL, Vogel-Walcutt JJ. A review of eye-tracking applications as tools for training. *Cognition Technol Work.* 2013;15(3):313-327. 10.1007/s10111-012-0234-7
28. Wilbanks BA, McMullan SP. A review of measuring the cognitive workload of electronic health records. *CIN: Computers, Informatics, Nursing.* 2018;36(12):579-588.
29. Hess EH, Polt JM. Pupil size in relation to mental activity during simple problem-solving. *Science.* 1964;143(3611):1190-1192.
30. Hyönä J, Tommola J, Alaja A-M. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology.* 1995;48(3):598-612.
31. Kahneman D, Beatty J. Pupil diameter and load on memory. *Science.* 1966;154(3756):1583-1585.
32. Van Orden KF, Limbert W, Makeig S, Jung T-P. Eye activity correlates of workload during a visuospatial memory task. *Hum Factors.* 2001;43(1):111-121.
33. Klingner J, Tversky B, Hanrahan P. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology.* 2011;48(3):323-332. 10.1111/j.1469-8986.2010.01069.x

34. Eckstein MK, Guerra-Carrillo B, Miller Singley AT, Bunge SA. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Dev Cogn Neurosci.* 2017;25: 69-91. https://doi.org/10.1016/j.dcn.2016.11.001

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

36. Holmqvist K, Andersson R. *Eye-tracking: A comprehensive guide to methods, paradigms and measures.* Lund, Sweden: Lund Eye-Tracking Research Institute; 2017.

37. Chen H, Dey A, Billinghurst M, Lindeman RW. Exploring pupil dilation in emotional virtual reality environments. 2017.

38. Meehan M, Razzaque S, Insko B, Whitton M, Brooks FP. Review of four studies on the use of physiological reaction as a measure of presence in stressfulvirtual environments. *Appl Psychophysiol Biofeedback.* 2005;30(3):239-258.

39. Paas FGWC. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics - a Cognitive-Load Approach. *J Educ Psychol.* 1992;84(4):429-434. 10.1037/0022-0663.84.4.429

40. Watson AB, Yellott JI. A unified formula for light-adapted pupil size. *Journal of Vision.* 2012; 12(10):12-12. 10.1167/12.10.12

41. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001

42. Hayes TR, Petrov AA. Mapping and correcting the influence of gaze position on pupil size measurements. *Behav Res Methods.* 2016;48(2):510-527. 10.3758/s13428-015-0588-x

Chapter 7

General Discussion

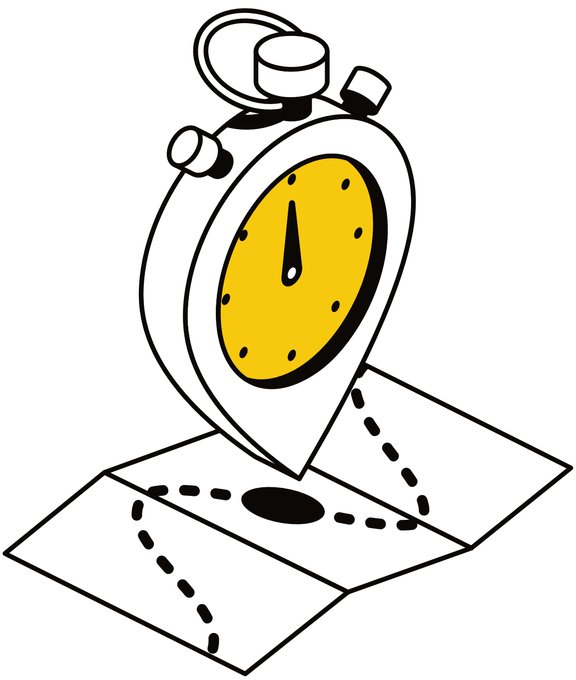
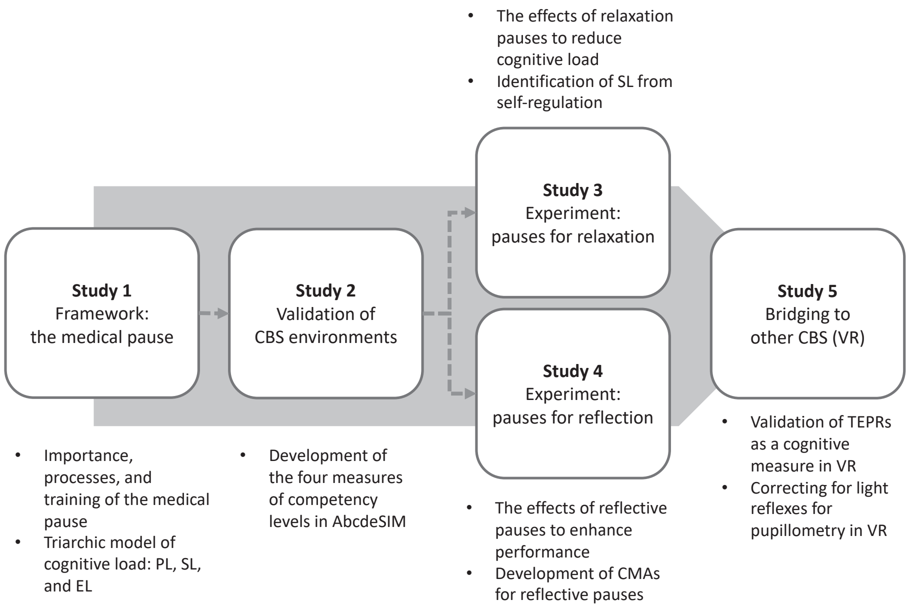


Figure 21 presents an overview of the five studies and their main findings reported in this PhD thesis. To deepen the understanding of pausing as a professional skill by investigating the pause effects theoretically and empirically, Study 1 provided a theoretical framework for the medical pause, while the following four studies empirically examined pause effects in computer-based simulation (CBS), based on this framework. After Study 2 validated the four measures and the task environment of the AbcdeSIM, Study 3 and 4 revealed the effects of pausing in terms of relaxation and reflection during pauses. In preparation for future studies, Study 5 validated pupillometry to measure cognitive load in VR environments.

Figure 21. Overview of the five studies and their main findings



In this final chapter, I describe the main findings from each study including a meta-analysis of their combined observations. Then, I discuss the theoretical and methodological contributions of the presented studies to the scientific literature as well as their limitations. Next, practical implications for educators and CBS designers follow. Finally, the main conclusion of all the work included in this thesis is presented.

1. Findings from this Thesis

Study 1 established the overall conceptual framework of the medical pause, by explicating its importance, constituent processes, and implementation in training programs. By using cognitive load theory, I presented a new triarchic model of primary load (PL, cognitive load caused by domain-specific performance), secondary load (SL, caused by domain-general processes such as self-regulation), and extraneous load (EL, caused by task-irrelevant processes). To maximize performance, the working memory should be neither overloaded nor underloaded by the total amount of cognitive load. The pausing skill was introduced as the professional skill that finds and maintains this balance through relaxation and reflection.

The identification of SL is particularly vital to critical fields such as medicine, since SL is closely related with "mindful" performance including self-monitoring and situation awareness,[1,2] which directly affects patient safety.[3,4] Moreover, the recognition of SL as a positive process that contributes to performance and learning broadens the perspective toward cognitive load. There have been similar studies in general education that emphasize the role of SL. For instance, Wolfgang and Kürschner stressed the importance of imposing cognitive load to enhance learning as it brings learners to the zone of proximal development.[5] Studies on desirable difficulties[6] and cognitive load caused by self-regulation[7,8] can also be interpreted along the same line. I argue that the conceptualization of the two types of loads that are associated with the cognitive and metacognitive processes relevant for learning (i.e., PL and SL) provides a useful framework for research in medical education, which unfolds in the following studies.

Study 2 laid the basis for empirical studies on the medical pause in CBS, by developing sensitive measures of performance and cognitive load in a CBS environment. First, I presented a model of problem-solving that represents the level of expertise by using theories from complex learning (e.g., 4C/ID[9]). Next, a task analysis with skill hierarchy description followed, given a CBS environment of AbcdeSIM,[10] a simulation game for emergency medicine. Eventually, multiple indicators of performance (i.e., systematicity, visual strategy, accuracy, speed) and cognitive load were discovered, and their validity was examined in a controlled experimental setup.

The findings of Study 2 emphasize the importance of theory-based cognitive task analysis that should precede the validation of measures in CBS. Theory-based task analysis allows for an integral assessment of multiple aspects of performance, which is more effective than one-dimensional evaluation.[11,12] This is particularly substantial for assessment in medicine where problem-solving processes contain extremely complex aspects (e.g., routine or non-routine, algorithmic or heuristic).[9,13] I presume the measures' validity and sensitivity found in this study stem from the well-developed task analysis and fitting operationalization to a specific task environment. Thanks to these validated measures, I discovered that students tend to experience cognitive overload while playing AbcdeSIM, which would benefit from pausing.

Study 3 showed how relaxation takes place during pauses and may help reduce cognitive overload. This study used an intense scenario in AbcdeSIM with unexpected advert events (i.e., epileptic seizure randomly attacks the patient), and a performance measure that represents how effectively learners counteract the events (i.e., intensity-care score). When pausing was available, participants showed both higher cognitive load and performance, regardless whether they actually used the pause function or not. I consider that this higher cognitive load indicates the existence of SL, confirming that self-monitoring to deploy pauses can produce extra cognitive load. This self-monitoring affected performance positively, resulting in higher intensity-care scores. Pausing temporarily reduced cognitive load, which demonstrates that students tend to use pauses for relaxation to cope with their heightened cognitive load.

Interestingly, given that pauses were available to be used freely, 50% of participants did not take any pauses. Moreover, compared to these participants, the other 50% who actually took pauses did not show any improvement in performance. The rare use of pausing has also been reported by some studies in general education.[14-16] It suggests that students neither are aware of the utility of pausing nor have skills to use pausing to contribute to performance. I reckon that these results from Study 3 confirm the argument from Study 1: Pausing is a professional skill that only experts with a developed clinical judgement and time management skills can use properly. To help novice learners use pausing productively, well-designed instructional support is necessary.

Study 4 demonstrated how reflection during pauses with the instructional support can optimize cognitive load and improve performance. I developed three types of cognitive and metacognitive aids (CMAs) to support the reflection: prompts, cues, and leading questions. Performance was measured in two categories: domain-specific performance (i.e., diagnosis and intervention) and domain-general performance (i.e., vigilance). Heightened cognitive load decreased and vigilance was improved. However, these effects emerged only in the later stage of the learning process, and the learners' domain-specific performance did not improve.

The key point of this study is the importance of well-designed reflection support. To date, there has been an intensive debate on how to define reflective pauses and whether they are beneficial or not.[17-19] I argue that there is not one type of beneficial reflective pause: How to design CMAs to make the pauses effective for given tasks is more important. Study 4 provides a good example of these CMAs. Reflection was prompted between part-tasks to maximize integration of constituent skills, and cues and leading questions were provided to facilitate reflection on performance. These CMAs appear to make the effects of reflective pauses prominent enough to be found in this study.

However, I find two limitations in the CMAs in Study 4: the latency of the effects, and the absence of feedback. Since the positive effects emerge only in the later stage of the learning process (i.e., latency), future research should investigate these effects by including an adaptation period or using longitudinal setups. I also suspect the absence of feedback prevented the domain-specific performance from improving, as feedback can provide points-of-improvement and schema-structuring.[9,20] Although I intentionally excluded the feedback process in this study to test the exact effects of reflection, future studies should examine the effects of feedback integrated with other CMAs.

Study 5 sets the stage for future studies in VR, by validating measures of cognitive load in the new environment. Although pupillometry has been reliably used as a measure of cognitive load in 2D CBS, its applicability to VR environments had not been tested before. Notoriously, pupil light reflexes caused by the VR display can confound task-evoked pupillary responses (TEPRs) as a cognitive load measure. I controlled for this confounding factor by modelling the light reflexes, inspired by methods from the human-computer interaction field. By using a controlled experimental design where task difficulty was manipulated as a single factor, I determined that TEPRs yield a valid cognitive load measure in VR training. To my knowledge, this study is the first to show this validity, paving the way for future studies exploring the pause effects in VR by using a reliable measure of cognitive load.

The leitmotif in this project is that pausing is not a skill that simply drops cognitive load and enhances performance just by taking arbitrary pauses. Availability of pausing can even impose extra cognitive load as it evokes other types of cognitive processes to regulate when and how to pause (i.e., self-regulation), as corroborated by Study 3 (the higher cognitive load in the pause-available condition) and Study 4 (the higher cognitive load in the reflective-pause condition in the earlier stage). However, this extra cognitive load is not necessarily detrimental if it is caused by metacognitive processes beneficial for performance and learning, as theoretically explained by Study 1 and empirically supported by Study 3 (enhanced performance in the pause-available condition). Moreover, the increased cognitive load can gradually be optimized as learners continue practicing the medical pause, given that they are guided by well-designed CMAs (the decrease in cognitive load in the later stage in Study 4). Once learners become more experienced in the long term, their cognitive load may be fully optimized as shown by the experts' lower cognitive load in Study 2.

Importantly, among diverse aspects of performance, the biggest beneficiary of pausing appears to relate to patient safety. In this project, two performance measures were enhanced by pausing: the intensity-care score (representing how well participants react to sudden advert events) in Study 3, and the vigilance on patient status in Study 4. I consider that the SL stimulated by the availability of pausing enhanced the intensity-care score in Study 3, while the visual cues provided during reflective pauses encouraged students to be attentive to the patient status in Study 4. This is a particularly meaningful finding for medical education as standards for patient safety become higher each year. Future studies in medical education could expand the research on this specific benefit of pausing, focusing on the effects of activation of SL and the provision of cues to support reflection.

2. Theoretical and Methodological Contributions

This project contributes to the literature in three ways: (1) deepened understanding of the medical pause in healthcare training, (2) advanced application of new constructs (e.g., the triarchic model) in educational psychology (i.e. cognitive load theory, 4C/ID), and (3) the development of innovative performance measures in CBS by using game logs and eye-tracking.

For medical education, this project provides new approaches to safety culture by introducing the medical pause. I focused on the basic nature of pausing and the cognitive mechanisms shared between formal timeouts with structured protocols in a team setting[21,22] and informal timeouts that can occur individually without formal protocols.[23,24] The medical pause, then, was conceptualized as a cognitive practice that allows for mindful and adaptive performance across this continuum of formal and informal timeouts. It connects key concepts in safety culture, such as the slowing down phenomenon identified by Moulton et al.[1,24,25] and the use of checklists by Pronovost et al.[26-28] This furthers the existing discussion on timeouts, and encourages practitioners to consider how to adopt a simple habit of pausing in workplaces. Moreover, the medical pause was identified as a professional skill that should be systematically taught in educational programs. This project also provided practical techniques to implement pausing in training programs, which serves as a translational education effort.

For educational psychology, my research made theoretical contribution to psychological and educational models, in particular, cognitive load theory (CLT) and complex learning (4C/ID). In the CLT community, there have been continuing discussions about types of cognitive load, such as intrinsic, germane, and extraneous load.[5,29] I argue that, while this traditional model is suitable to stable task environments where learning is the main focus, it is less applicable to explain cognition in dynamic and critical task environments where both learning and performance should be controlled by metacognitive processes. Thus, I established a new triarchic model of cognitive load with PL, SL, and EL as distinct constructs. This model is particularly useful to explain the interconnection between cognitive load, self-regulation, and learning processes. It contributes to the existing discussion on interplay between cognitive load and self-regulation.[7,8] Also, the terminology presented in this project may facilitate explicit communication between researchers in cognitive psychology and education, promoting future studies on the interplay between cognitive and metacognitive processes in highly dynamic task environments.

According to 4C/ID, the metacognitive processes to control performance and learning require support to function and develop properly. I introduced concepts from 4C/ID, such as *second-order scaffolding* and *part-task sequencing*, to produce a good example of CMAs for taking reflective pauses. A part-task sequencing method of *forward chaining with snowballing* was used to prompt pauses at the points where the interconnection between different cognitive skills is maximized. The *guided discovery method* was used to help learners start from what they already know and develop their schemas by answering leading questions. This project demonstrated how key concepts from 4C/ID can be applied to provide instructional support for metacognitive processes in dynamic task environments. Based on this example, future studies can further investigate different forms of second-order scaffolding, by studying the effects of CMAs on learning and developing instructional designs for metacognitive processes such as reflective pausing.

From a methodological viewpoint, this project has shown an advanced application of diverse data sets (e.g., game logs, eye-tracking) to measure cognitive processes in performance and learning. By combining these data sets, I developed multiple indicators of performance, such as systematicity, accuracy, speed, and cognitive load. Throughout the project, I demonstrated (1) how machine learning techniques can be used to measure important aspects of performance such as systematicity, (2) how synchronizing game logs and eye-tracking data can yield valuable information, and (3) how pupillometry can be refined to be a reliable measure of cognitive load by employing methods from human-computer interaction. This provides good examples of the development of informative and reliable measures, facilitating future studies in performance assessment in diverse CBS task environments.

Additionally, this project showed the explicit strengths of eye-tracking for research in CBS: It can (1) reveal hidden cognitive features (e.g., expertise, cognitive load, vigilance) that subjective ratings cannot capture, (2) provide support for reflection by recording visual behaviors and presenting them as a cue, and (3) be applied to diverse format of CBS such as 2D serious games or 3D VR, as these CBS task environments heavily depend on visual stimuli. My project showed that future studies using such task environments could benefit from these specific advantages of eye-tracking.

3. Limitations

There are several limitations in this project. First, I did not consider social factors that could influence the effects of pausing (e.g., personality of practitioners, social pressure).[24] For example, social pressure to maintain a communicative mode during a surgery could affect the initiation of pausing. This social effect of pausing can be studied further in the field of team learning in medical practice.

Second, although Study 1 mentioned the possibility of negative effects of pausing, it was not specifically investigated in the empirical studies. In extreme situations where expediency should be prioritized (e.g., emergency cesarean section), the benefit of pausing can be doubted. In teamwork, an individual's pausing could deteriorate collective performance. These negative aspects of pausing can be further studied by considering more factors and consequences of pausing.

Third, this project tested the pause effects only in simulation-based task environments. Future studies could extend the research to real-task environments or the actual workplace. Fourth and lastly, since this project focused on the short-term effects of pausing, future studies could test its long-term effects using longitudinal study setups.

4. Practical Implications

Table 16 presents instructional methods and their applications in three different phases of training: before, during, and after the training. In existing simulation-based training programs, educators can readily implement pausing following these methods. As preparation before the training, learners should be informed of the importance of and strategies for taking pauses. CMAs should be designed based on cognitive task analysis. The degree of support by supervisors should be negotiated based on learners' competency levels. The terminology from the medical pause can be used to facilitate this communication.

Table 16. Instructional methods and their applications for training pausing skills in different phases

Phase	Instructional methods	Application
Preparation before training	· Shared understanding on the medical pause · Design of CMAs · Negotiation for degree of scaffolding, based on competency assessment	· Supervisors inform learners of why and how to pause · Learners anticipate the scenario and plan when and how to pause, while supervisors guide them with examples · For novice learners, supervisors should take full control in scaffolding · Educators design CMAs for given scenarios based on task analysis
During training in classrooms	· Provision of CMAs (prompts, cues, and leading questions) to support pausing · Integration of planning and improvising of pauses	(When using physical simulation) · Supervisors facilitate reflection during pauses by using CMAs · Supervisors share control over improvising pauses with learners. · Cognitive overload is managed by removing resource of EL (When using CBS) · Learners share CBS experience with peers in classrooms through a big screen · Learners discuss with peers and receive feedback during pauses · Cognitive overload is managed by lowering fidelity of CBS
Remote (online) learning after the class	· Repetitive practice through online access to CBS · Implementation of CMAs in CBS	· Learners practice the scenario and the medical pause at home, supported by CMAs implemented in the CBS · CBS designers develop CBS with a pause function and CMAs

During training, pauses are initiated through support of the CMAs, while additional pauses can be improvised guided by the supervisors. When using CBS in a classroom, a large screen shares what learners are playing in the CBS. During the pauses, students reflect on their performance and discuss with peers on how to improve their performance. After each pause, they can take turns leading the scenario to apply the newly found solutions in real time. As cognitive load can be too high due to the activation of SL, in case of highly demanding simulation tasks, other types of cognitive load should be reduced, for instance, by lowering the fidelity levels of the simulation (e.g., removing sound effects from the simulation).

After the training in the classroom, students can repeatedly practice the scenario and the pausing strategies at home by using online access to the CBS. Students' performance data can be stored to track their development, and replayed for later reflection. In consultation with educators and researchers, CBS designers could implement CMAs that automatically support training of pause skills, to make the remote online learning effective.

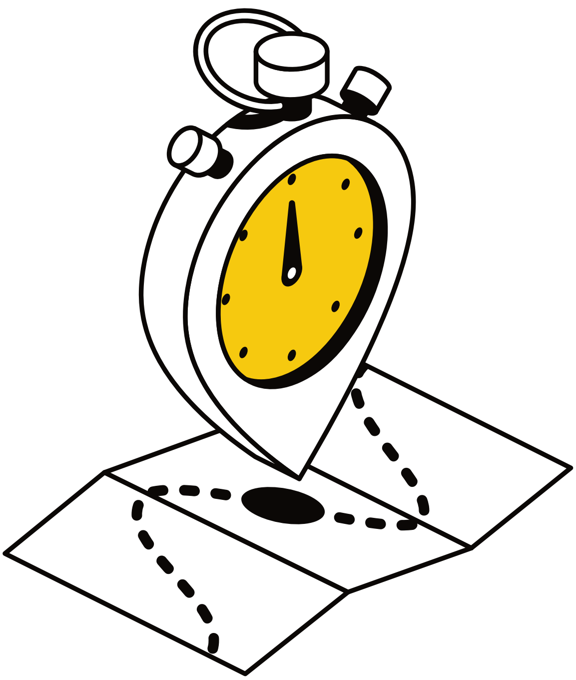
5. Conclusion

Strategies for safety training often fail: Professionals hardly conform to timeout protocols and trainees disregard prompts to pause for reassessment of safety measures. As "culture eats strategy every time",[30] to shift the focus of discussion from strategy to culture, discussion on the medical pause should be nurtured through research, communication, and education. In highly dynamic and critical task environments, healthcare practitioners are either overloaded or underloaded, leading to undesirable consequences. Although it can sound contradictory or uncomfortable to healthcare professionals, studies have proven that the medical pause can bring about fundamental changes. Through training supported by enriched cognitive and metacognitive support, the medical pause can be learned as an integral part of clinical expertise. By mastering the medical pause, healthcare professionals profit from its benefits, knowing when and how to pause and so improve the patient safety.

Reference

1. Moulton CA, Epstein RM. Self-monitoring in surgical practice: Slowing down when you should. In: Fry H. KR, ed. *Surgical Education*. Springer, Dordrecht; 2011:169-182.
2. Endsley MR. Situation awareness. *Handbook of human factors and ergonomics.* 2006;3:528-542.
3. Blandford A, William Wong BL. Situation awareness in emergency medical dispatch. 2004;61(4): 421-452. 10.1016/j.ijhcs.2003.12.012
4. Tien G, Atkins MS, Zheng B, Swindells C. Measuring situation awareness of surgeons in laparoscopic training. Paper presented at: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications; March, 2010.
5. Schnotz W, Kürschner C. A reconsideration of cognitive load theory. *Educ Psychol Rev.* 2007; 19(4):469-508.
6. Bjork RA, Bjork EL. Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition.* 2020;9(4):475-479.
7. Seufert T. The interplay between self-regulation in learning and cognitive load. *Educational Research Review.* 2018;24:116-129. 10.1016/j.edurev.2018.03.004
8. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001
9. Van Merriënboer JJ, Kirschner PA. *Ten steps to complex learning: A systematic approach to four-component instructional design.* New York: Routledge; 2018.
10. *AbcdeSIM* [computer program]. Rotterdam: VirtualMedSchool; 2012.
11. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education.* 1996;1(3):227-233. 10.1007/BF00162920
12. Dankbaar ME, Stegers-Jager KM, Baarveld F, et al. Assessing the assessment in emergency care training. *PLoS One.* 2014;9(12). 10.1371/journal.pone.0114663
13. Van Merriënboer JJ. Perspectives on problem solving and instruction. *Computers & Education.* 2013;64:153-160. 10.1016/j.compedu.2012.11.025
14. Hasler BS, Kersten B, Sweller J. Learner control, cognitive load and instructional animation. *Appl Cogn Psychol.* 2007;21(6):713-729. 10.1002/acp.1345
15. Kettanurak V, Ramamurthy K, Haseman WD. User attitude as a mediator of learning performance improvement in an interactive multimedia environment: an empirical investigation of the degree of interactivity and learning styles. *Int J Hum Comput Stud.* 2001;54(4):541-583. 10.1006/ijhc.2001.0457
16. Moreno R, Valdez A. Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Etr&D-Educational Technology Research and Development.* 2005;53(3):35-45. Doi 10.1007/Bf02504796
17. Kori K, Pedaste M, Leijen Ä, Mäeots M. Supporting reflection in technology-enhanced learning. *Educational Research Review.* 2014;11:45-55. 10.1016/j.edurev.2013.11.003

18. Taub M, Azevedo R, Bradbury AE, Mudrick NV. Self-regulation and reflection during game-based learning. *Handbook of Game-Based Learning.* 2020:239-262.
19. Yang X, Liu Y. Supporting Students' Reflection in Game-Based Science Learning: A Literature Review. In: Li R. CSKS, Iwasaki C., Kwok LF., Kageto M. , ed. *Lecture Notes in Computer Science.* Springer International Publishing; 2021:119-131.
20. Hattie J, Timperley H. The Power of Feedback. *Review of Educational Research.* 2007;*77*(1): 81-112. 10.3102/003465430298487
21. Mahajan RP. The WHO surgical checklist. *Elsevier BV.* 2011; 25(2): 161-168. 10.1016/j.bpa.2011.02.002
22. Schmutz JB, Eppich WJ. Promoting Learning and Patient Care Through Shared Reflection. *Acad Med.* 2017;92(11):1555-1563. 10.1097/acm.0000000000001688
23. St-Martin L, Patel P, Gallinger J, Moulton CA. Teaching the slowing-down moments of operative judgment. *Surgical Clinics.* 2012;92(1):125-135. 10.1016/j.suc.2011.12.001
24. Moulton CA, Regehr G, Lingard L, Merritt C, Macrae H. 'Slowing Down When You Should': Initiators and Influences of the Transition from the Routine to the Effortful. *J Gastrointest Surg.* 2010; 14(6):1019-1026. 10.1007/s11605-010-1178-y
25. Moulton CA, Regehr G, Mylopoulos M, Macrae HM. Slowing Down When You Should: A New Model of Expert Judgment. *Acad Med.* 2007; 82(Suppl): S109-S116. 10.1097/acm.0b013e3181405a76
26. Pronovost PJ, Vohr E. *Safe patients, smart hospitals: how one doctor's checklist can help us change health care from the inside out.* New York, NY: Plume; 2011.
27. Hales BM, Pronovost PJ. The checklist—a tool for error management and performance improvement. *J Crit Care.* 2006;21(3):231-235. 10.1016/j.jcrc.2006.06.002
28. Pronovost PJ, Bo-Linn GW. Preventing Patient Harms Through Systems of Care. *JAMA.* 2012; 308(8):769. 10.1001/jama.2012.9537
29. Kalyuga S. Cognitive load theory: How many types of load does it really need? *Educ Psychol Rev.* 2011;23(1):1-19.
30. Melnyk BM. Culture Eats Strategy Every Time: What Works in Building and Sustaining an Evidence-Based Practice Culture in Healthcare Systems. *Worldviews Evid Based Nurs.* 2016;13(2): 99-101. 10.1111/wvn.12161



Summary

This PhD thesis aims to foster the understanding of the medical pause as a professional skill that should be taught in educational programs, by investigating the effects of pausing both theoretically and empirically. In *Chapter 1*, General Introduction, I review previous studies on pausing skills in medical education and education in general. To explain cognitive mechanisms that constitute pausing skills, cognitive load theory is introduced. Relaxation and reflection are stressed as the two major cognitive processes that make pausing effective for increasing performance and learning. Computer-based simulation (CBS) is illustrated as a promising environment for pausing skills for both educational and research purposes. This introductory chapter presents an overview of the five studies that build up this thesis. Study 1 provides a theoretical framework of the medical pause. After Study 2 validates performance measures and the task environment of a CBS, namely, a serious game. Study 3 and 4 investigate the two processes during pauses, that is, relaxation and reflection. Finally, Study 5 tests the generalizability of the findings in another type of CBS environment, Virtual Reality (VR).

Chapter 2 establishes a systematic conceptualization of the medical pause, focusing on its importance, processes, and implementation in training programs. By employing insights from educational sciences and cognitive psychology, I first identify pausing as an important skill to interrupt negative momentum and bolster learning. Subsequently, I categorize constituent cognitive processes for pausing skills into two phases: the decision-making phase (determining when and how to take pauses) and the executive phase (applying relaxation or reflection during pauses). I present a new model that describes how relaxation and reflection during pauses can optimize cognitive load in performance. Several strategies to implement pause training in medical curricula are proposed: intertwining pause training with training of primary skills, providing second-order scaffolding through shared control, and employing auxiliary tools such as CBS with a pause function.

Chapter 3 aims to validates performance measures and the task environment of a serious game for emergency medicine (i.e., AbcdeSIM) by testing whether the measures

developed in this study can predict different levels of prior knowledge. Based on theories of complex-skill acquisition (e.g., 4C/ID), I derive four performance aspects that prior knowledge may affect: (1) systematicity in approach, (2) accuracy in visual attention and motor reactions, (3) speed in performance, and (4) cognitive load. The measures are developed to represent these aspects by using machine learning, game-log analysis, and eye-tracking.

Participants were 24 medical professionals (experts, with high prior knowledge) and 22 medical students (novices, with low prior knowledge). After pre-training, they all played one scenario, during which game logs and eye movements were collected. A cognitive-load questionnaire ensued. During game play, experts demonstrated a more systematic approach, higher accuracy in visual selection and motor reaction, and a higher performance speed than novices. Their reported levels of cognitive load were lower. These results indicate that prior knowledge has a substantial impact on performance in AbcdeSIM, opening up the possibility of using the measures for performance assessment.

Chapter 4 investigates how pausing affects performance and cognitive load in intense situations in AbcdeSIM. On the assumption that allowing pauses and actually taking pauses are two different kinds, the effects of these two constructs are tested respectively. Medical students (N = 70) were randomly assigned to one of two conditions: simulation with (n = 40) and without (n = 30) the option to take pauses. All participants played the same two scenarios, during which game logs and eye-tracking data were recorded.

Overall, both cognitive load and performance were higher in the condition with pauses than in the one without pauses. The act of pausing, however, temporarily lowered cognitive load, especially during intense moments. Two different manifestations of the pause effect were identified: (1) by stimulating additional cognitive and metacognitive processes, pauses increased overall cognitive load; and (2) through relaxation, the act of pausing temporarily decreased heightened cognitive load. Consequently, these results suggest that in order to enhance students' performance and learning it is important to encourage them to utilize the different effects of pausing depending on the given situation.

Chapter 5 examines the effects of reflective pauses on performance, given that instructional support for reflection is provided. By using concepts from complex learning, I propose how to design cognitive and metacognitive aids (CMAs) to support reflection processes during pauses. Assuming reflective pauses with CMAs help to optimize cognitive load and allow for restructuring of mental models, I examine their effects on four aspects of performance and learning: cognitive load, domain-specific performance, domain-general performance, and the structure of cognitive schemas.

Medical students (N = 72) performed tasks in the AbcdeSIM task environment, in two conditions: reflection condition (n = 36) where reflection was prompted during pauses, and control condition (n = 36) without such prompts. The effects of reflective pauses only emerged in the later stage of the learning process, while no significant effects were identified in the early stage. Cognitive load decreased and the domain-general aspects of performance improved. However, domain-specific performance aspects and schema structure did not improve, probably due to lacking feedback during reflection. These results suggest that theory-based support design can make in-action reflection more effective, demonstrating that reflective pauses can enhance performance, but an adaptation period is required.

Chapter 6 explores whether the indicators found in the previous studies can be applied to VR environments. Although pupillometry is well-known as a reliable technique to measure cognitive load in 2D environments, its applicability to 3D VR environments had not been validated yet. Specifically, the VR display causes light reflexes that confound task-evoked pupillary responses (TEPRs). Through this pilot study, I validate whether task difficulty can predict cognitive load as measured by TEPRs corrected for the light reflex and if these TEPRs correlate with cognitive load self-ratings and performance.

Fourteen students in health sciences performed observation tasks in two conditions: difficult versus easy tasks, whilst watching a VR scenario in home health care. Then, a cognitive load self-rating ensued. I used a VR system with a built-in eye-tracker and a photosensor installed to assess pupil diameter and light intensity during the scenario. Employing a method from the human-computer interaction field, I determined TEPRs by modeling the pupil light reflexes using a baseline. As predicted, the difficult task caused significantly larger TEPRs than the easy task. Only in the difficult task condition did TEPRs positively correlate with the performance measures. These results suggest that TEPRs are valid measures of cognitive load in VR training when corrected for the light reflex. It opens up possibilities to use real-time cognitive load for assessment and instructional design for VR training.

Chapter 7 is the General Discussion which brings the findings from all studies together. By connecting these findings, it discusses how the studies in this thesis contributed to the investigation of the pause effects and deepened the understanding of the medical pause. Throughout the five studies, the medical pause is identified as a complex professional skill that should be understood through interplay between cognitive load, self-regulation, and learning processes. Among diverse aspects of performance that pausing can enhance, the biggest beneficiary of pausing appears to relate to patient safety.

Theoretical and methodological contributions of this project are identified in three fields: (1) healthcare training (2) educational psychology (i.e. cognitive load theory, 4C/ID), and (3) performance assessment in CBS. This project provides new approaches to safety culture by introducing the medical pause, and facilitates the existing discussion on the interplay between cognitive load and self-regulation by establishing a new triarchic model of cognitive load with PL, SL, and EL. By using game logs and eye-tracking, it presents an advanced application of diverse data sets to measure cognitive processes in performance and learning.

Limitations of the studies are illustrated to guide future researchers to unfold the research on the medical pause. Also, practical implications for educators and CBS designers are discussed by suggesting instructional applications. For instance, the degree of support by supervisors should be negotiated based on learners' competency levels before the training. During the training, pauses are initiated through support of the CMAs, while additional pauses can be improvised guided by the supervisors. After the training, students can repeatedly practice the scenario and the pausing strategies at home by using online access to the CBS.

Samenvatting

Dit proefschrift beoogt het begrip voor medische pauzes te bevorderen als een professionele vaardigheid die in onderwijsprogramma's onderwezen zou moeten worden, door de effecten van pauzeren zowel theoretisch als empirisch te onderzoeken. In *hoofdstuk 1*, Algemene inleiding, geef ik een overzicht van eerdere studies over vaardigheden in het geven van pauzes in het medisch onderwijs en het onderwijs in het algemeen. Om de cognitieve mechanismen van pauzerende vaardigheden te verklaren, wordt de cognitieve belastingstheorie geïntroduceerd. Ontspanning en reflectie worden benadrukt als de twee belangrijkste cognitieve processen die pauzeren effectief maken voor het verbeteren van prestaties en leren. Computergebaseerde simulatie (CBS) wordt gegeven als een veelbelovende omgeving voor pauzerende vaardigheden voor zowel onderwijs- als onderzoeksdoeleinden. Dit inleidende hoofdstuk geeft een overzicht van de vijf studies waaruit dit proefschrift is opgebouwd. Studie 1 geeft een theoretisch kader over het medisch pauzeren. Hierna, in studie 2, worden prestatiemetingen en de taakomgeving van een CBS, namelijk een serious game, gevalideerd. Studie 3 en 4 onderzoeken de twee processen tijdens pauzes, namelijk ontspanning en reflectie. Tot slot test studie 5 de generaliseerbaarheid van de bevindingen in een ander type CBS-omgeving, namelijk Virtual Reality (VR).

Hoofdstuk 2 geeft een systematische conceptualisering van de medische pauze, met aandacht voor het belang, de processen en de implementatie ervan in opleidingsprogramma's. Door gebruik te maken van inzichten uit de onderwijswetenschappen en de cognitieve psychologie, identificeer ik eerst het nemen van pauzes als een belangrijke vaardigheid om negatief momentum te onderbreken en het leren te versterken. Vervolgens categoriseer ik de cognitieve processen die nodig zijn om te kunnen pauzeren in twee fasen: de besluitvormingsfase (bepalen wanneer en hoe te pauzeren) en de uitvoerende fase (ontspanning of reflectie toepassen tijdens pauzes). Ik presenteer een nieuw model dat beschrijft hoe ontspanning en reflectie tijdens pauzes de cognitieve belasting in de uitvoering kunnen optimaliseren. Verschillende strategieën om pauzetraining te implementeren in medische curricula worden voorgesteld: het verweven van pauzetraining met training van primaire vaardigheden, het bieden van second-order scaffolding door middel van gedeelde controle, en het inzetten van hulpmiddelen zoals CBS met een pauzefunctie.

Hoofdstuk 3 heeft tot doel de prestatiemaatstaven en de taakomgeving van een serieuze game voor spoedeisende geneeskunde (i.e., AbcdeSIM) te valideren door te testen of de in deze studie ontwikkelde maatstaven verschillende niveaus van voorkennis kunnen voorspellen. Gebaseerd op theorieën over het verwerven van complexe vaardigheden (bv. 4C/ID), leid ik vier prestatieaspecten af die door voorkennis beïnvloed kunnen worden: (1) systematiek in aanpak, (2) nauwkeurigheid in visuele aandacht en motorische reacties, (3) snelheid in uitvoering, en (4) cognitieve belasting. De maatstaven worden ontwikkeld om deze aspecten weer te geven door middel van machine learning, game-log analyse, en eye-tracking.

De deelnemers waren 24 medische professionals (experts, met hoge voorkennis) en 22 medische studenten (nieuwelingen, met lage voorkennis). Na de pre-training speelden ze allemaal één scenario, waarbij game-logs en oogbewegingen werden verzameld. Daarna werd een vragenlijst over cognitieve belasting afgenomen. Tijdens het spel vertoonden experts een meer systematische aanpak, een hogere nauwkeurigheid in visuele selectie en motorische reactie, en een hogere uitvoeringssnelheid dan nieuwelingen. De door hen vermelde niveaus van cognitieve belasting waren lager. Deze resultaten geven aan dat voorkennis een substantiële invloed heeft op prestaties in AbcdeSIM, wat de mogelijkheid opent tot het gebruiken van de maatstaven voor prestatiebeoordeling.

Hoofdstuk 4 onderzoekt hoe pauzeren prestaties en cognitieve belasting beïnvloedt in intense situaties in AbcdeSIM. Uitgaande van de veronderstelling dat het toestaan van pauzes en het daadwerkelijk nemen van pauzes twee verschillende soorten zijn, worden de effecten van deze twee opstellingen getest. Medisch studenten ($N = 70$) werden willekeurig toegewezen aan een van de twee condities: simulatie met- ($n = 40$) en zonder ($n = 30$) de mogelijkheid om pauzes te nemen. Alle deelnemers speelden dezelfde twee scenario's, waarbij game logs en eye-tracking data werden geregistreerd.

Over het geheel genomen waren zowel de cognitieve belasting als de prestaties hoger in de conditie met pauzes dan in de conditie zonder pauzes. De handeling van het pauzeren verlaagde echter tijdelijk de cognitieve belasting, vooral tijdens intense momenten. Twee verschillende manifestaties van het pauze-effect werden geïdentificeerd: (1) door extra cognitieve en metacognitieve processen te stimuleren, verhoogden pauzes de totale cognitieve belasting; en (2) door ontspanning, verminderde de handeling van het pauzeren tijdelijk de verhoogde cognitieve belasting. Deze resultaten suggereren dat, om de prestaties en het leren van studenten te verbeteren, het belangrijk is hen aan te moedigen de verschillende effecten van pauzeren te gebruiken, afhankelijk van de gegeven situatie.

Hoofdstuk 5 onderzoekt de effecten van reflectieve pauzes op de prestaties, gegeven dat instructieve ondersteuning voor reflectie wordt geboden. Door gebruik te maken van concepten uit complex learning, stel ik voor hoe cognitieve en metacognitieve hulpmiddelen (CMAs) ontworpen kunnen worden om reflectieprocessen tijdens pauzes te ondersteunen. Ervan uitgaande dat reflectieve pauzes met CMA's helpen om cognitieve belasting te optimaliseren en herstructurering van mentale modellen mogelijk te maken, onderzoek ik hun effecten op vier aspecten van prestatie en leren: cognitieve belasting, domeinspecifieke prestatie, domein-generalere prestatie, en de structuur van cognitieve schema's.

Studenten geneeskunde (N = 72) voerden taken uit in de AbcdeSIM taakomgeving, in twee condities: reflectieconditie (n = 36), waar reflectie werd gevraagd tijdens pauzes, en controleconditie (n = 36) zonder dergelijke prompts. De effecten van reflectiepauzes kwamen pas naar voren in de latere fase van het leerproces, terwijl er geen significante effecten werden gevonden in de vroege fase. De cognitieve belasting nam af en de domeinspecifieke aspecten van de prestatie verbeterden. Domeinspecifieke prestatieaspecten en schemastructuur verbeterden echter niet, waarschijnlijk als gevolg van het ontbreken van feedback tijdens de reflectie. Deze resultaten suggereren dat een op theorie gebaseerd ondersteuningsontwerp in-actie reflectie effectiever kan maken, en tonen aan dat reflectieve pauzes de prestaties kunnen verbeteren, maar dat een aanpassingsperiode nodig is.

In hoofdstuk 6 wordt onderzocht of de indicatoren die in de eerdere studies zijn gevonden, kunnen worden toegepast op VR omgevingen. Hoewel pupillometrie bekend staat als een betrouwbare techniek om cognitieve belasting te meten in 2D omgevingen, was de toepasbaarheid op 3D VR omgevingen nog niet gevalideerd. Het VR beeldscherm veroorzaakt lichtreflexen die taak-geëvokeerde pupilresponsen (TEPRs) verstoren. Met deze pilootstudie wil ik nagaan of de moeilijkheidsgraad van de taak cognitieve belasting kan voorspellen aan de hand van TEPRs gecorrigeerd voor de lichtreflex en of deze TEPRs correleren met cognitieve belasting en prestatie.

Veertien studenten gezondheidswetenschappen voerden observatietaken uit in twee condities: moeilijke versus makkelijke taken, terwijl ze naar een VR-scenario in de thuisgezondheidszorg keken. Daarna volgde een zelfbeoordeling van de cognitieve belasting. Ik gebruikte een VR-systeem met een ingebouwde eye-tracker en een fotosensor om de pupildiameter en de lichtintensiteit tijdens het scenario te beoordelen. Gebruikmakend van een methode uit het human-computer interaction veld, bepaalde ik TEPRs door de pupil licht reflexen te modelleren met behulp van een baseline. Zoals voorspeld, veroorzaakte de moeilijke taak aanzienlijk grotere TEPRs dan de gemakkelijke taak. Alleen in de moeilijke taakconditie correleerden TEPRs positief met de prestatiemetingen. Deze resultaten suggereren dat TEPRs geldige maatstaven zijn voor cognitieve belasting in VR training wanneer gecorrigeerd wordt voor de lichtreflex. Het opent mogelijkheden om real-time cognitieve belasting te gebruiken voor beoordeling en instructieontwerp voor VR training.

Hoofdstuk 7 is de Algemene Discussie waarin de bevindingen uit alle studies worden samengebracht. Door deze bevindingen met elkaar te verbinden, wordt besproken hoe de studies in dit proefschrift hebben bijgedragen aan het onderzoek naar de pauze-effecten en het begrip van de medische pauze hebben verdiept. Doorheen de vijf studies wordt de medische pauze geïdentificeerd als een complexe professionele vaardigheid die begrepen moet worden door de wisselwerking tussen cognitieve belasting, zelfregulatie, en leerprocessen. Onder de verschillende aspecten van de prestaties die pauzeren kan verbeteren, lijkt de grootste begunstiging van pauzeren betrekking te hebben op de veiligheid van de patiënt.

Theoretische en methodologische bijdragen van dit project worden geïdentificeerd op drie gebieden: (1) opleiding in de gezondheidszorg (2) onderwijspsychologie (d.w.z. cognitieve belastingstheorie, 4C/ID), en (3) prestatiebeoordeling in het CBS. Dit project biedt nieuwe benaderingen voor veiligheidscultuur door de medische pauze te introduceren, en vergemakkelijkt de bestaande discussie over de wisselwerking tussen cognitieve belasting en zelfregulatie door een nieuw triarchisch model van cognitieve belasting met PL, SL, en EL op te stellen. Door gebruik te maken van game logs en eye-tracking, wordt een geavanceerde toepassing van diverse datasets gepresenteerd om cognitieve processen in prestaties en leren te meten.

Beperkingen van de studies worden geïllustreerd om toekomstige onderzoekers een leidraad te geven bij het uitbouwen van het onderzoek naar de medische pauze. Ook worden praktische implicaties voor opleiders en CBS-ontwerpers besproken door instructieve toepassingen voor te stellen. Bijvoorbeeld, de mate van ondersteuning door supervisors zou moeten worden onderhandeld op basis van het competentieniveau van de lerenden vóór de opleiding. Tijdens de training worden pauzes geïnitieerd door ondersteuning van de CMA's, terwijl extra pauzes kunnen worden geïmproviseerd onder leiding van de supervisors. Na de training kunnen de cursisten het scenario en de pauzestrategieën thuis herhaaldelijk oefenen door gebruik te maken van de online toegang tot het CBS.

Impact Chapter

The medical pause is about strengthening safety culture. This project deepens the understanding of the medical pause, by producing conceptual models and evidence of its effects on learning in computer-based simulation (CBS). The findings reinforce the discussions on safety culture in healthcare by advancing research, education, and technology. In the following, this contribution is described in three parts: scientific impact (safety culture and medical education, educational science), societal impact (educators, CBS designers, and technology industry), and dissemination activities.

1. Scientific Impact

Safety Culture and Medical Education

Patient safety is one of the major topics in healthcare research. This project provides new approaches to research on safety by connecting existing key concepts such as timeouts,[1,2] checklists,[3-5] and slowing down phenomena,[6-8] and introducing an integral concept of the medical pause. The cognitive mechanisms shared among the previous studies are identified and consolidated based on cognitive psychology, which expands the focus of discussion from strategy (e.g., timeout protocols) to culture (e.g., shared understanding of pausing). This moves the literature on safety culture one step further and allows future studies to build upon the literature through the newly developed understanding and terminology.

For medical education and training, it proposes the medical pause as a professional skill that should be taught already from early phases in training programs. To serve as a translational educational effort, it provides practical techniques to implement pausing in training programs as well as examples of application. This provision may be of value to educators seeking to design better training programs that foster the habit of reflective practice. They can start by introducing simple applications such as checklists that prompt learners to plan and improvise pauses on their own.

Educational Science

For educational psychology, this project further develops the existing discussion on interplay between cognitive load and self-regulation.[9,10] Furthermore, it establishes a new triarchic model of cognitive load with primary, secondary, and extraneous load. This model is particularly powerful in explaining the interconnection between cognitive load, self-regulation, and learning processes. The concepts and terminology presented in this project (e.g., pausing skills, planning and improvising, relaxation and reflection) can facilitate explicit communication between researchers in education, promoting future studies on the interplay between cognitive load and self-regulation.

Research on CBS and gamification also benefits from this project. I developed multiple indicators for performance assessment based on educational theories by using e and eye-tracking data. This provides good examples of the development of reliable measures, facilitating future studies in performance assessment in CBS task environments. Especially, eye-tracking has shown exceptional strengths for research in CBS: It can (1) reveal hidden cognitive processes (e.g., cognitive load, vigilance) that subjective ratings cannot capture, (2) provide support for reflection by recording visual behaviors and presenting them as a cue, and (3) be applied to diverse formats of CBS such as 2D serious games or VR, as these CBS task environments heavily depend on visual stimuli. Future studies on learning in CBS can refer to the methods used in this project to utilize these potentials of eye-tracking.

2. Societal Impact

Educators

Educators and trainers can immediately implement pausing in existing programs. As prompting learners to plan and reflect has already been used as a teaching method in many training programs, pausing to reflect can be readily integrated. As I suggested in Study 1, supervisors should assess the learners' competency beforehand, then discuss how much control on planning or improvising of pausing can be granted to the learners. The terminology from the medical pause can be used to facilitate this discussion.

The use of CBS and eye-tracking to train pausing skills can be integrated in current educational settings, such as problem-based learning and blended learning. In a classroom, a large screen shares what learners are playing in the CBS, while a computer with an eye-tracker is saving the data. In small groups, students discuss with peers on when and how to pause to improve learning and performance. During the pauses, they take turns leading the scenario to apply the newly found solutions in real time. At home, students can repeatedly practice what they learned in the classroom by using online access. Students' performance data can be stored to track their development and re-played for later reflection.

Computer-Based Simulation (CBS) Designers

Although many CBS systems already have a pause function available, CBS designers are not sufficiently aware of the importance of providing support to make pauses contribute to learning. This project presents a good example of this support, namely, cognitive and metacognitive aids (CMAs) with prompts, cues, and leading questions. CBS designers can use this example when implementing a pause function in their systems. CMAs should be designed to fit the learning goals through a task analysis, which requires active communication between CBS designers, educators, and researchers.

Importantly, when implementing a pause function with CMAs, the extra cognitive load that might be imposed on learners should be taken into account. Although the extra cognitive load could contribute to performance (i.e., secondary load), if the total cognitive load exceeds the learners' working memory capacity, it can be detrimental to their learning processes. Thus, especially in case of highly demanding tasks, the designers should consider reducing other types of cognitive load (e.g., lowering physical fidelity of the simulation) to optimize working memory and learning processes.

Technology Industry

As this NWO-funded project involved game developers such as IJsfontein (www.ijsfontein.nl) and Ranj (www.ranj.com), and the game provider of VirtualMedSchool (www.virtualmedschool.com), the research findings have been shared with them throughout the project, and informed them of educational demands of effective CBS design. Since this project is expanded to VR, more organizations (e.g., VR hardware and software companies) can be involved. Such collaboration will allow for more diverse technology in VR to be used in education in the future. For instance, increasing the use of haptic sensors, voice control for communication skills, integration of AI for more realistic virtual patients, multiplayer VR (multiple learners work together in the same scenarios), and adaptive systems that tailor themselves to learners' progress.

While eye-tracking software for 2D stimuli has been well developed by several companies (e.g., SMI, Tobii), the software for VR learning environments is in a very early stage. To develop a functioning software that fits my research purpose, I have collaborated with Vizard (www.worldviz.com), a VR software company in the USA. We developed an eye-tracking package where eye-tracking data can be integrated with data from photo sensors, which is now commercially available. In our ongoing collaboration, we keep developing the software to advance other types of eye-tracking measures (e.g., gaze allocation in VR space), allowing for in-depth research in 3D space and VR training in the future.

3. Dissemination Activities

The research outcomes have been shared by publishing in peer-reviewed journals. Study 1, 2, and 3 were published with open access, while Study 4 and 5 are currently undergoing the review process. Paper presentations and symposia have been arranged at national and international conferences: DSSH (Dutch Society for Simulation in Healthcare) 2019, Association for Medical Education in Europe (AMEE) 2020, ORD (Onderwijs Research Dagen) 2021, EARLI (European Association for Research on Learning and Instruction) 2019 and 2021. In ICLTC (International Cognitive Load Theory Conference) 2018, 2019 and 2021, the application of cognitive load theory to medical education was elaborated. In ECEM (European Conference on Eye Movements) 2019, the methodological aspects of eye-tracking were focused on. Local departmental meetings and interviews were also used to disseminate the findings from my PhD project: Lunch Lecture in Faculty of Health, Medicine, and Life Sciences, Lunch session in Skillslab, Online Learning & Instruction Round Table at Open University in the Netherlands, and Expert Interview for Serious Game in Medical Education at Leiden University.

The research outcomes are also shared through the SHE eye-tracking lab (https://youtu.be/MO9PHBSjOzk) and local networks in Maastricht University such as the FHML VR/AR workgroup, TRI-SIM, and FHML Simulation Day. During this project, IRESIM (International Community for Research on Simulation, https://www.medsimresearch.org) was initiated. It is an international community where researchers, designers, and educators share up-to-date ideas about simulation in medical education. The research findings were shared at the regular webinar of IRESIM, and nurtured through discussion between the members.

References

1. Backster A, Teo A, Swift M, Polk HC, Harken AH. Transforming the Surgical "Time-Out" Into a Comprehensive "Preparatory Pause". *J Card Surg.* 2007;22(5):410-416. 10.1111/j.1540-8191.2007.00435.x
2. Altpeter T, Luckhardt K, Lewis JN, Harken AH, Polk Jr HC. Expanded surgical time out: a key to real-time data collection and quality improvement. *J Am Coll Surg.* 2007;204(4):527-532. 10.1016/j.jamcollsurg.2007.01.009
3. Pronovost PJ, Vohr E. *Safe patients, smart hospitals: how one doctor's checklist can help us change health care from the inside out.* New York, NY: Plume; 2011.
4. Hales BM, Pronovost PJ. The checklist—a tool for error management and performance improvement. *J Crit Care.* 2006;21(3):231-235. 10.1016/j.jcrc.2006.06.002
5. Pronovost PJ, Bo-Linn GW. Preventing Patient Harms Through Systems of Care. *JAMA.* 2012; 308(8):769. 10.1001/jama.2012.9537
6. Moulton CA, Epstein RM. Self-monitoring in surgical practice: Slowing down when you should. In: Fry H. KR, ed. *Surgical Education.* Springer, Dordrecht; 2011:169-182.
7. Moulton CA, Regehr G, Lingard L, Merritt C, Macrae H. 'Slowing Down When You Should': Initiators and Influences of the Transition from the Routine to the Effortful. *J Gastrointest Surg.* 2010; 14(6):1019-1026. 10.1007/s11605-010-1178-y
8. Moulton CA, Regehr G, Mylopoulos M, Macrae HM. Slowing Down When You Should: A New Model of Expert Judgment. *Acad Med.* 2007; 82(Suppl): S109-S116. 10.1097/acm.0b013e3181405a76
9. De Bruin ABH, Van Merriënboer JJG. Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction.* 2017;51:1-9. 10.1016/j.learninstruc.2017.06.001
10. Seufert T. The interplay between self-regulation in learning and cognitive load. *Educational Research Review.* 2018;24:116-129. 10.1016/j.edurev.2018.03.004

Acknowledgements

Jeroen (vM), Jeroen (D), and Halszka. Thank you for being my triumvirate that breaks a lance for my PhD. Jeroen (vM), you would never know how much you inspired me. You are not just a great supervisor but an exceptionally humane mentor. You kindled my love for science to fully enjoy it. I always remember your saying, "Joy, I think you have no limitations", which has become one of my life-quotations. During the last 4 years, I have been extremely happy to expand my horizons without limitations, thanks to your being my lighthouse.

Jeroen (D), my steady and brilliant signpost. While the other Jeroen encourages me go through the high seas, you always protected me providing great anchors. I could grow as an outstanding data scientist with various tools in my hands, thanks to your guidance and encouragement. Also, you have shown me how one can be both kind in mind and sharp in thinking. Our deep discussions about music, culture, and languages have been always a great fun and intellectual enjoyment.

Halszka, my great role model as a shining female researcher. I always admire your leaderships in various networks and practical advice which create breakthroughs against heavy obstacles. You inspired me to be keen in science and passionate for my dream. You have been always standing there whenever I reach out for help, fully supporting me to become a professional researcher.

The wonderful colleagues at the IGBL research team. Many thanks to Mary, Tjitske, Walter, Stephanie, Ronald, and Evert. What a great team we formed together across industry, research, and education. We were a unique combination that inspired the game industry and research with our expertise and passion.

The medical pause team across borders. Thanks to Adam, Daniel, Heather, and John, how great that we combined our experience in both educational psychology and clinical work in different cultural contexts! We learned a lot from each other, exchanging the effective and thoughtful support flying across Atlantic.

Many thanks to Paul and Fred, the legendary brain-eye-skin-heart team. I'm grateful for your great inspiration, and admire the high-level humor code that runs through every meeting. I still giggle whenever recalling the punching kangaroos in Australia, and the wine festival during Zooms.

My old colleagues in Lund, Agneta and Magnus. I always feel warm and supported for your continuous encouragement from the very time I left for the Netherlands. Richard, my first and best mentor! The eye-tracking book from you has been always on my desk, playing with me, throughout my PhD.

The VR team in Regensburg. Irina, Angelika, and Daniel, thank you for your invaluable help and expertise. I'm inspired by not only your expertise but also the great family-like teamwork. Many thanks to Ryan and Myra, for the great expertise for review studies. I learned a lot and am grateful for your kind support.

TRISIM colleagues for whom everyday is Simulation Day festival. Thank you, Uli, Femke, Patricia, Dewa, and Nynke, for helping my research to go smooth and sharing your thoughts. Special thanks to Géraldine, for her great knowledge as both healthcare professionals and teacher, and accurate feedback and support for ABCDE research. Jimmy, Maryam, and Gabrielle, thanks for your invaluable expertise and support for my VR research and funding applications.

The phenomenal colleagues at SHE. Anique, Diana, and Pim, thank you for being so caring and supportive, with your nurturing leadership. Renée and Pascal, thank you very much for your inspiring comments during SHE academy and conferences. Special remark for the truly amazing support by Nicky, Hennie, Eveline, Lilian, Audrey, Ruud, Maurice, Angelique, and Joost! Your support is not just efficient, but also, I could always enjoy fun and intimate conversations with you.

Many thanks to my brilliant PhD companions, Steph, Luo, Carolin, Andrea, Lorette, Emmaline, Lianne, Felicitas, Juliët, Anne, Michelle, AniqueSuzanne, Sannes, Erdem, Hector, Fury, Derk, Serge, Fatameh, Koos, Jolien, Tiuri, Cindy, HQ, Fengyuan, Guy, and Tiuri.

I enjoyed all types of discussions with you at every corner, corridors, Banditos, on the bikes, and under trees.

Nico-chan, the wise, thank you for the amazing FICS service where thousands of hedgehogs kindly listen to my complex thoughts and give me surprising wisdom. Jason, ddoneun Sson, you are the most honorable Dutch in some particular Korean's history. I love our endless conversation about languages and history, and admire your deep knowledge and kindness.

Der sagenhafte Walprofessor der Walsee! The penguin pupil in the red basket always applauds your fabulous fountain. Pippi-nim, my polite little bird or the rabbit in the cup. I'll always be there to give a nod of approval for your dreams whatsoever. My Um-mi and Ah-ppi, the love between us is unconditional. Whatever the future brings, my deepest heart will be with you, nonstop hugging.

About the Author

Joy Yeonjoo Lee started her PhD in the School of Health Professions Education at Maastricht University, the Netherlands. Her research has focused on improving instructional design in medical education, through emerging technology and data science. In particular, she has active interest in instructional design theories, eye tracking methodology, VR/AR training environments, and application of data science to learning analytics.

Prior to the PhD, she received a BA in business administration from Yonsei University, South Korea. She has work experience in IT fields as a data analyst and project manager. She earned a MSc in cognitive science, and has done her master's project from Lund University, Sweden. She did her internship in Educational Technology Group at Lund University, working for eye-tracking research projects and data analysis.

List of Publications

Lee JY, Donkers J, Jarodzka H, Sellenraad G, Faber TJE, & van Merriënboer JJ. The Reflective Pause in Simulation Training: The Effects of In-action Reflection on Cognitive Load and Performance. (submitted)

Lee JY, de Jong N, Donkers J, Jarodzka, H, van Merriënboer JJ. Measuring Cognitive Load in Virtual Reality Training via Pupillometry. (submitted)

Holmqvist K, Orbom SL, Hooge ITC, Niehorster DC, Alexander RG, Andersson R, Benjamins JS, Blignaut P, Brouwer AM, Chuang LL, Dalrymple KA, Drieghe D, Dunn MJ, Ettinger U, Fiedler S, Foulsham T, van der Geest JN, Hanson DW, Hutton S, Kasneci E, Kingstone A, Knox PC, Kok EM, Lee H, **Lee JY,** Leppänen JM, Macknik SM, Majaranta P, Martinez-Conde S, Nuthmann A, Nyström M, Orquin JL, Otero-Millan J, Park SY, Popelka S, Proudlock F, Renkewitz F, Roorda AJ, Schulte-Mecklenbeck MS, Sharif B, Shic F, Shovman M, Thomas MG, Venrooij W, Zemblys R, Hessels RS. Eye tracking: Empirical Foundations for Reporting Guidelines. *Behavior Research Methods*. 2021 Nov 29.

Lee JY, Szulewski A, Young JQ, Donkers J, Jarodzka H, van Merriënboer JJ. The medical pause: Importance, processes and training. *Medical Education*. 2021 Oct 1;55(10):1152-60.
* This article is also published as Apple Podcast: https://podcasts.apple.com/us/podcast/the-medical-pause-importance-processes-and-training/id784455563?i=1000538584056

Ayres P, **Lee JY**, Paas F, van Merriënboer JJ. The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. Frontiers in Psychology. 2021 Sep 10;12.

Tärning B, **Lee YJ**, Andersson R, Månsson K, Gulz A, Haake M. Assessing the black box of feedback neglect in a digital educational game for elementary school. *Journal of the Learning Sciences*. 2020 Oct 19;29(4-5):511-49.

Lee JY, Donkers J, Jarodzka H, Sellenraad G, Van Merriënboer JJ. Different effects of pausing on cognitive load in a medical simulation game. *Computers in Human Behavior*. 2020 Sep 1;110:106385.

Lee JY, Donkers J, Jarodzka H, Van Merriënboer JJ. How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Computers in Human Behavior*. 2019 Oct 1;99:268-77.

SHE Dissertation 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

Ilgen, J. (15-12-2021) Comfort with uncertainty in medical professionals. An exploration of how clinicians experience and manage dynamic problems in practice

Schut, S. (9-12-2021) The Burden of Proof - Agency and Accountability in Programmatic Assessment'

Hui, L. (6-12-2021) 'Fostering Self-Regulated Learning: the Role of Perceived Mental Effort'

Meeuwissen, S. (12-11-2021) 'Team learning at work. Getting the best out of interdisciplinary teacher teams and leaders'

Nguyen Thi, V.A. (02-11-2021) Motivating and educating health professionals to work in less attractive specialties: Findings from experiences of Vietnam

Martens, S. (15-10-2021) Building student-staff partnerships in higher education

Lestari, E. (05-10-2021) INTERPROFESSIONAL EDUCATION Lessons from Indonesia

Atherley, A. (27-09-2021) Beyond the struggles: Using social-developmental lenses on the transition to clinical training

Schillings, M. (06-07-2021) Talking about feedback: Face-to-face peer dialogue about written feedback

Wilbur, K. (05-07-2021) NO WHERE | NOW HERE: Context and Competency Expectations in Workplace-Based Training

Bendermacher, G. (02-07-2021) Navigating from Quality Management to Quality Culture

Ahmed Khan, R. (29-06-2021) Assessing curriculum viability in Undergraduate Medical Education

Chim, H.Q. (30-03-2021) Physical Activity Behavior and Learning in Higher Education

Dominguez, L.C. (23-02-2021) Persistence in surgical training: The role of job crafting and leadership

Bindels, E. (22-02-2021) DOING WELL, GETTING BETTER; Facilitating physicians' reflection on their professional performance

Iqbal, Z. (15-12-2020) All stakeholders matter in faculty development: Designing entrustable professional activities for small group facilitation

Tran, QT. (09-12-2020) Nationwide implementation of medical skills training laboratories in a developing country: studies from Vietnam

Pacifico, J. (30-11-2020) Making the Implicit Explicit: Uncovering the Role of the Conceptions of Teaching and Learning and the Perceptions of the Learning Climate in Postgraduate Medical Training.

Nishigori, H. (17-11-2020) Why do doctors work for patients? Medical professionalism in the era of neoliberalism

Oudkerk Pool, A. (06-11-2020) Competency-based portfolio assessment – Unraveling stakeholder perspectives and assessment practices

Geel van, K. (05-11-2020) Lifelong learning in radiology: all eyes on visual expertise

Stammen, L. (16-10-2020) Pursuing - High-Value, Cost-Conscious Care - The Role of Medical Education

Meulen van der, M. (15-10-2020) Assessment of physicians' professional performance using questionnaire-based tools

Matsuyama, Y. (05-10-2020) Contextual attributes fostering self-regulated learning in a teacher-centered culture: learner's professional identity formation is a trigger

Rovers, S. (16-09-2020) Growing knowledge, supporting students' self-regulation in problem-based learning

Bourgeois-Law, G. (03-09-2020) Conceptualizations of remediation for practicing physicians

Giuliani, M. (19-05-2020) A Critical Review of Global Curriculum Development, Content and Implementation in Oncology

Schreurs, S. (20-03-2020) Selection for medical school; the quest for validity

Schumacher, D. (19-03-2020) Resident Sensitive Quality Measures: Defining the Future of Patient-Focused Assessment

Sehlbach, C. (21-02-2020) To be continued.... Supporting physicians' lifelong learning

ICO Dissertation Series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, Radboud University Nijmegen, University of Amsterdam, University of Antwerp, University of Ghent, KU Leuven, Université Catholique de Louvain, University of Groningen, University of Twente, Utrecht University, Vrije Universiteit Amsterdam, and Wageningen University, and formerly Tilburg University (until 2002). List update February 10, 2022 (the list will be updated every year in January)

458. Vaessen, B.E. (08-01-2021) Students' perceptions of assessment and student learning in higher education courses. Eindhoven: Eindhoven University of Technology
459. Maureen, I.Y. (15-01-2021) Story time in early childhood education: designing storytelling activities to enhance (digital) literacy development. Enschede: University of Twente
460. Van Alten, D.C.D. (19-03-2021) Flipped learning in secondary education history classrooms: what are the effects and what is the role of self-regulated learning. Utrecht: Utrecht University
461. Gestsdóttir, S.M. (08-03-2021) Observing history teaching: historical thinking and reasoning in the upper secondary classroom. Amsterdam: University of Amsterdam
462. Chim, H.Q. (30-03-2021) Physical Activity Behavior and Learning in Higher Education. Maastricht: Maastricht University
463. Krijnen, E. (15-04-2021)Family literacy in context: Exploring the compatibility of a family literacy program with children's homes and schools. Rotterdam Erasmus University Rotterdam

464. Stolte, M. (07-05-2021) (In)attention for creativity: Unraveling the neural and cognitive aspects of (mathematical) creativity in children. Utrecht: Utrecht University
465. Rathé, S. (12-05-2021) Focusing on numbers – An investigation of the role of children's spontaneous focusing on Arabic number symbols in early mathematical development. Leuven: KU Leuven
466. Theelen, H. (12-05-2021) Looking around in the classroom. Developing preservice teachers' interpersonal competence with classroom simulations. Wageningen: Wageningen University
467. De Jong, L.A.H. (20-05-2021) Teacher professional learning and collaboration in secondary schools. Leiden: Leiden University
468. Sincer, I. (20-05-2021) Diverse Schools, Diverse Citizens? Teaching and learning citizenship in schools with varying student populations. Rotterdam: Erasmus University Rotterdam
469. Slijkhuis, E.G.J. (20-05-2021) Fostering active citizenship in young adulthood. Groningen: University of Groningen
470. Groothuijsen-Vrancken, S.E.A. (02-06-2021) Quality and impact of practice-oriented educational research. Utrecht: Utrecht University
471. Hingstman, M. (07-06-2021) Supporting struggling students: prevention and early intervention with Success for All. Groningen: University of Groningen
472. Gerdes, J. (14-06-2021) All inclusive? Collaboration between teachers, parents and child support workers for inclusive education in prevocational schools. Amsterdam: Vrije Universiteit Amsterdam
473. Bai, H. (18-06-2021) Divergent thinking in young children. Utrecht: Utrecht University
474. Wijnker, W. (23-06-2021) The Unseen Potential of Film for Learning: Film's Interest Raising Mechanisms Explained in Secondary Science and Math. Utrecht: Utrecht University
475. Brummer, L. (24-09-2021). Unrooting the illusion of one-size-fits-all feedback in digital learning environments. Groningen: University of Groningen
476. Veldman, M.A. (01-07-21) Better together, social outcomes of cooperative learning in the first grades of primary education. Groningen: University of Groningen
477. Wang, J. (06-07-2021) Technology integration in education: Policy plans, teacher practices, and student outcomes. Leiden: Leiden University
478. Zhang, X. (06-07-2021) Teachers' teaching and learning motivation in China. Leiden: Leiden University

479. Poort, I.C. (02-09-2021) Prepared to engage? Factors that promote university students' engagement in intercultural group work. Groningen: University of Groningen
480. Guo, P. (07-09-2021) Online project-based higher education Student collaboration and outcomes. Leiden: Leiden University
481. Jin, X. (21-09-2021) Peer feedback in teacher professional development. Leiden: Leiden University
482. Atherley, E.N. (27-09-2021) Beyond the struggles: Using social-developmental lenses on the transition to clinical training. Maastricht: Maastricht University
483. Martens, S.E. (15-10-2021) Building student-staff partnerships in higher education. Maastricht: Maastricht University
484. Ovbiagbonhia, R. (08-11-2021) Learning to innovate: how to foster innovation competence in students of Built Environment at universities of applied science. Wageningen: Wageningen University
485. Van den Boom-Muilenburg , S.N. (11-11-2021) The role of school leadership in schools that sustainably work on school improvement with professional learning communities. Enschede: University of Twente
486. Sachishal, M.S.M. (11-11-2021) Science interest - Conceptualizing the construct and testing its predictive effects on current and future behavior. Amsterdam: University of Amsterdam
487. Meeuwissen, S.N.E. (12-11-2021) Team learning at work: Getting the best out of interdisciplinary teacher teams and leaders. Maastricht: Maastricht University
488. Keijzer-Groot, A.F.J.M. (18-11-2021) Vocational identity of at-risk youth – Tailoring to support career chances. Leiden: Leiden University
489. Wolthuis, F. (25-11-2021) Professional development in practice. Exploring how lesson study unfolds in schools through the lens of organizational routines. Groningen: University of Groningen
490. Akkermans-Rutgers, M. (06-12-2021) Raising readers. Stimulating home-based parental literacy involvement in the first, second, and third grade. Groningen: University of Groningen
491. Hui, L. (06-12-2021) Fostering Self-Regulated Learning: The Role of Perceived Mental Effort. Maastricht: Maastricht university
492. Jansen, D. (08-12-2021) Shadow education in the Netherlands: The position of shadow education in the educational landscape and students' school careers. Amsterdam: University of Amsterdam
493. Kamphorst, F. (15-12-2021) Introducing Special Relativity in Secondary Education. Utrecht: Utrecht University

494. Eshuis, E.H. (17-12-2021) Powering Up Collaboration and Knowledge Monitoring: Reflection-Based Support for 21st-Century Skills in Secondary Vocational Technical Education. Enschede: University of Twente