Quantifying Learner Expertise Using Unobtrusive Measures of Cognitive Load During Training

Jeffrey M. Beaubien, Ph.D.[1], Rachel L. Elkin, MD[2], David O. Kessler[2], MD, Todd P. Chang, MD[3], Nathaniel Damaghi, B.A.[2], John Feeney, Ph.D.[1], William Noah DePriest, B.S.[1]

Aptima, Inc.[1], Columbia University[2], Children's Hospital[3]

Woburn, MA[1], New York, NY[2], Los Angeles, CA[3]

jbeaubien@aptima.com, re2357@cumc.columbia.edu, dk2592@cumc.columbia.edu, dr.toddchang@gmail.com, nd2215@cumc.columbia.edu, jfeeney@aptima.com, wdepriest@aptima.com

ABSTRACT

The construct of cognitive load (CL) is rooted in the dual-process theory of decision making, which postulates two distinct types of cognitive processes operating largely, but not completely, in parallel (Evans, 2003; Evans & Stanovich, 2013; Van Merriënboer & Sweller, 2010). One of these decision processes, "Type 1," is extremely fast, makes minimal demands on working memory, and operates by associatively comparing the current situation to one's corpus of accumulated prior experiences from long-term memory. Type 1 decision skills are consistent with the rapid and near-automatic recognition-primed decision making (RPD) approach used by domain experts. By comparison, "Type 2" decision processes involve conscious deliberation and explicit mental calculations, thereby placing heavy demands on working memory. Type 2 decision skills are consistent with the slow and effortful decision making approach used by domain novices (Kahneman & Klein, 2009). The purpose of the current study was to unobtrusively measure the CL of physicians using wireless, commercial off-the-shelf (COTS) neurophysiological monitors. The participants included a mixed sample of pediatricians (6 novices, 6 experts) who performed four different Virtual Reality (VR) simulation scenarios (2 clinical scenarios x 2 levels of distraction). After each scenario, the participants completed the NASA-Task Load Index (Hart & Staveland, 1988) self-reported measure of workload. A linear mixed model (LMM) revealed a significant main effect of expertise (experts had lower mean CL scores than novices), as well as a significant expertise-by-clinical scenario interaction. Similar, but not identical, effects were found using the TLX scores. Additional results, implications, and lessons learned are presented.

ABOUT THE AUTHORS

Dr. Jeffrey M. Beaubien is a Distinguished Principal Scientist in Aptima's Learning and Training Systems division. For the past 20 years, his work has focused on training and assessing leadership, teamwork, and decision making skills. His research has been sponsored by the U.S. Navy, the U.S. Army, the U.S. Air Force, and the Telemedicine and Advanced Technologies Research Center, among others. Dr. Beaubien holds a Ph.D. in Industrial and Organizational Psychology from George Mason University, a M.A. in Industrial and Organizational Psychology from the University of New Haven, and a B.A. in Psychology from the University of Rhode Island.

Dr. Rachel L. Elkin is a Pediatric Emergency Medicine Fellow at Columbia University Irving Medical Center. Her clinical and research interests focus on the intersection of simulation, medical education, and technology in healthcare. She holds an MD with special qualifications in Biomedical Research from Cleveland Clinic Lerner College of Medicine, as well as an M.S. in Clinical Research.

Dr. David O. Kessler is the Vice Chair of Innovations and Strategic Initiatives for the Department of Emergency Medicine at Columbia University Vagelos College of Physicians and Surgeons. He is co-director and a founder of the International Network for Simulation-Based Pediatric Innovation, Research, and Education (INSPIRE) that aims to improve the lives of children through healthcare simulation science. Dr. Kessler is widely published and has led a variety of multi-centered and grant-funded studies to investigate the best ways to leverage technology for improved care. His current research interests include artificial intelligence, virtual reality, natural language processing, and the integration of leading-edge technologies to optimize healthcare education and care delivery.

Dr. Todd P. Chang is the Divisional Director for Research and Scholarship for the Division of Emergency Medicine at Children's Hospital Los Angeles. He has been Principal Investigator and Co-Investigator on a variety of grant-funded and multi-center educational technology research studies that explore the best practices of gamification, serious games, and virtual reality in training healthcare providers, with significant peer-reviewed publications and conference proceedings. Dr. Chang has had successful collaboration with simulation companies including Laerdal, BreakAway Games, Oculus from FaceBook, A.I.Solve, and Bioflight VR.

Mr. Nathaniel Damaghi is the Information Technology (IT) Manager for the Department of Emergency Medicine at Columbia University Irving Medical Center. He holds a B.A. in Information Systems from the New Jersey Institute of Technology and is currently enrolled in the Executive M.S. in Technology Management program at Columbia University. His current research interests include technology in healthcare, technology in automotive racing, augmented reality, virtual reality, and radio-frequency identification (RFID).

Dr. John Feeney is a Principal Research Engineer at Aptima, Inc., who specializes in the design and development of enhanced decision support systems which leverage machine learning and artificial intelligence technologies over a range of domains including occupational safety, healthcare, cyber operations, and intelligence analysis. Dr. Feeney holds a Ph.D. in Applied Experimental Psychology from the Catholic University of America, an M.S. in Software Engineering from National University, and a B.A. in Computer Science from the State University of New York at Oswego.

Mr. William Noah DePriest is a Senior Software Engineer in Aptima's Performance Assessment and Augmentation Division. A substantial amount of Noah's work involves designing and developing software for human-system environments, including the integration of physiological sensors with task simulators so that human performance data can be collected, stored, processed, and visualized both in real-time and post hoc. He also works closely with scientists and engineers to integrate advanced analytics and algorithms into software architectures to provide real-time assessments of operator states. He holds a B.S. in Computer Engineering from Ohio Northern University.

Quantifying Learner Expertise Using Unobtrusive Measures of Cognitive Load During Training

Jeffrey M. Beaubien, Ph.D.[1], Rachel L. Elkin, MD[2], David O. Kessler, MD[2], Todd P. Chang, MD[3], Nathaniel Damaghi, B.A.[2], John Feeney, Ph.D.[1], William Noah DePriest, B.S.[1]

Aptima, Inc.[1], Columbia University[2], Children's Hospital[3]

Woburn, MA[1], New York, NY[2], Los Angeles, CA[3]

jbeaubien@aptima.com, re2357@cumc.columbia.edu, dk2592@cumc.columbia.edu, dr.toddchang@gmail.com, nd2215@cumc.columbia.edu, jfeeney@aptima.com, wdepriest@aptima.com

BACKGROUND

The construct of Cognitive Load (CL) is rooted in the dual-process theory of decision making, which postulates that there are two types of cognitive processes operating largely, but not completely, in parallel (Evans, 2003; Evans & Stanovich, 2013; Van Merriënboer & Sweller, 2010). One of these processes, "Type 1," operates entirely at the unconscious level. Type 1 decision making is extremely fast, makes minimal demands on working memory, and operates in part by associatively comparing the current situation to one's corpus of accumulated prior experiences from long-term memory. All humans engage in a considerable amount of Type 1 processing in their day-to-day lives, such as when reading, driving, and identifying everyday objects. In addition, experts have well-developed Type 1 decision skills in their specific domain of expertise (Kahneman & Klein, 2009). By comparison, "Type 2" decision processes operate entirely at the conscious level. Type 2 decision making is considerably slower, places heavy demands on working memory, and requires explicit mental calculations. It is akin to the slow and deliberate decision making approach used by domain novices, as well as the slow and deliberate approach used by domain experts when facing novel problems or situations for which their expertise has not fully prepared them (Kahneman & Klein, 2009). One way to ascertain a learner's level of expertise is to measure their CL as they perform the task (Dan & Reiner, 2017). With increasing expertise, CL will decrease because the learner no longer needs to consciously monitor their task performance. Instead, the learner relies on information chunking (Chase & Simon, 1973; Gobet et al., 2001) and muscle memory (Gray, 2017) to perform the task automatically.

As learners become more expert-like, there will also be predictable changes in their task performance. Specifically, the task performance of domain novices is slow and effortful. Lacking any direct, first-hand task experience, novices perform the task in a series of discrete steps and do so following textbook descriptions. Their task movements are often "jerky" rather than "smooth," and they often err by omitting critical task steps or by performing steps in incorrect order. Because they must consciously monitor their task performance to avoid making errors, their mental workload is further heightened (Dreyfus & Dreyfus, 1980). With additional training, learners reach the "competent" stage of development. At this stage, their task performance becomes somewhat faster and less error prone. They perform the task steps in a more logical sequence, and there is a reduction in the number of redundant task steps performed. With even further training, learners reach the "proficient" and "expert" stages, respectively. During these stages, task performance becomes increasingly quicker, smoother, more accurate, and more efficient. The performer may also intentionally deviate from textbook descriptions, for example by chunking multiple task steps to achieve increases in speed or economy of motion. As the task becomes increasingly automatic, fewer cognitive processes are required to monitor task performance in real-time (Dreyfus & Dreyfus, 1980). In fact, consciously monitoring one's performance impedes the performance of domain experts (Wulf, McNevin, & Shea, 2001). Ideally, both measures of CL and task performance should be used to assess expertise, because relying on measures of CL alone can be misleading. For example, low levels of CL coupled with high levels of task performance would indicate an expert who is operating using Type 1 decision processes. By comparison, low levels of CL coupled with low levels of task performance would indicate a learner who is likely not taking the task seriously.

In a previous study (Beaubien, Wiggins, & DePriest, 2019), the current authors developed real-time, unobtrusive measures of individual and team CL using wireless commercial-off-the-shelf (COTS) neurophysiological monitors and custom developed CL models and classifiers (Durkee, Geyer, Pappada, Ortiz, & Galster, 2013; Pappada et al., 2016). The streaming CL measures were then tested with a sample of 15 multi-disciplinary healthcare teams who

performed a mix of clinical scenarios using a lifelike patient manikin in a simulated operating room. Mean levels of CL were computed separately for each team during each training scenario. The results suggest that as the teams' mean CL increased: their technical proficiency decreased (r = -.58, p. < .01); the number of appropriate clinical behaviors decreased (r = -.17, *ns*); the number of inappropriate clinical behaviors increased (r = .23, *ns*); and their self-reported workload increased (r = .27, p. <. 05). While not all of these results were statistically significant, all were in the hypothesized direction. However, no main effect of scenario difficulty was observed. In retrospect, the scenario difficulty manipulations were not as distinct as intended. For one set of simulation scenarios, mean CL values for the "medium" difficulty condition were extremely similar to the "low" difficulty condition. For another set of simulation scenarios, mean CL values for the "medium" difficulty condition were extremely similar to the "high" difficulty condition. Additionally, because of recruiting limitations, all of the participants were medical residents (novices). Because all participants were using Type 2 decision processes, it was not possible to empirically test for the effect of differences in expertise. Based on the lessons learned from that prior study, the streaming CL models and classifiers were revised to better control for the effects of physical motion by de-noising the neurophysiological signals. We then designed and conducted the current validation study to better evaluate the effects of expertise and scenario difficulty on CL and task performance.

PROCEDURE

The purpose of the current study was to evaluate the extent to which the revised measures of individual CL were working as intended. Specifically, we sought to answer the following three research questions: *To what extent do the unobtrusive CL measures differentiate between known groups of domain novices and experts?*, *To what extent do they differentiate among training scenarios of differing clinical content areas and distraction levels?*[1], and *To what extent do they correlate with the participants' own self-reports?*

Method

After providing informed consent, each participant was outfitted with a BioRadio 150 (Great Lakes Neurotechnologies, Cleveland, OH) wireless neurophysiological sensor, which was connected to a wet electrode electroencephalogram (EEG) cap and an electrocardiogram (ECG) strap which was worn around the chest. Next, the learner donned an Oculus Rift S (Facebook Technologies LLC, Menlo Park, CA) Head Mounted Display (HMD) (see Figure 1). The experimental apparatus was then tested to ensure that the neurophysiological signals were recorded accurately. During this time, each participant's CL was baselined by having them describe what they saw on the VR system's home screen, and by talking about whatever else they desired. However, the participants did not perform any experimental tasks at this time.

Figure 1. A member of the research team demonstrating the experimental apparatus.

[1] To maintain the participants' interest, all of the training scenarios were set to the highest level of task difficulty. The amount of external distraction was manipulated by varying the number and salience of device alarms, Public Address (PA) system announcements, and crying parents in the treatment bay to increase the task's extraneous cognitive load.

Next, each participant completed a brief, interactive tutorial on how to use the VR simulator. The tutorial addressed tasks such as navigating through the environment, interacting with the patient, selecting objects, performing diagnostic and therapeutic procedures, and the like. Finally, each participant completed a series of four simulation scenarios (Chang, Beshay, Hollinger, & Sherman, 2019), which were presented in counterbalanced blocks of content (anaphylactic shock vs. prolonged seizure). The order of scenario presentation within each block (low vs. high distraction) also was counterbalanced. After each training scenario, the participants retrospectively self-reported their level of workload before starting the next one. At the end of the experiment, the participants completed a brief demographic questionnaire that inquired about their domain expertise and familiarity with VR and gaming technologies. Finally, the participants were debriefed and thanked for their time. The entire data collection process lasted about 90 minutes.

Participants

The participants included a mix of novice (n = 6) non-board-certified pediatric resident physicians and expert (n = 6) board-certified or eligible pediatric emergency medicine physicians. The novices included three pediatricians who were in their first year of medical residency, and three who were in their second year. The experts included two fellows and four attending physicians with at least two more years of emergency experience than the novice group. Each participant completed a brief questionnaire that inquired about the number of times they had previously treated patients with anaphylactic shock (anaphylaxis) and prolonged seizure (status epilepticus), conditions that are relatively common in pediatric emergency medicine. Each question was answered using a 4-point rating scale with the following anchors: 1 = "never," 2 = "between 1-5 times", 3 = "between 6-10 times," and 4 = "greater than 10 times." The two questionnaire items were highly correlated (r = .98, p. <. 01). Each of the domain experts had treated greater than ten cases of each condition, while all of the domain novices had treated five or fewer cases of each. As expected, there was a statistically significant correlation between domain expertise and the treatment of patients with anaphylactic shock ($\chi^2_{(2)}$ = 12.00, p. < .01) and prolonged seizure ($\chi^2_{(2)}$ = 12.00, p. < .01), respectively. Together, these two questionnaire items confirm the domain experts' higher level of emergency resuscitation experience.

The questionnaire also inquired about the number of times that participants used VR and video game technology, respectively. Each question was answered using a 4-point rating scale with the following anchors: 1 = "never," 2 = "1-2 times weekly," 3 = "3-4 times weekly," and 4 = "5 or more times weekly." The two questionnaire items were uncorrelated (r = .29, *ns*). Regarding VR technology, 75% of the entire sample reported having never used the technology. The remaining 25% responded having used it 1-2 times weekly. There was no relationship between VR technology use and domain expertise ($\chi^2_{(1)}$ = .44, *ns*) or sex ($\chi^2_{(1)}$ = .44, *ns*); therefore, the results are not reported separately. Regarding gaming, all of the participants reported using video games 1-2 times weekly or less. However, there was a marginal relationship between expertise and gaming use ($\chi^2_{(1)}$ = 3.09, p. = .08). Specifically, among the experts, 83% reported never using games, and 17% reported using games 1-2 times weekly. By comparison, only 33% of the novices reported never using games, while 67% reported using games 1-2 times weekly. However, there was no relationship between gaming use and sex ($\chi^2_{(1)}$ =.34, *ns*). Taken together, the questionnaire results suggest that while both groups had limited familiarity using gaming and VR technologies, the novices had slightly more familiarity with gaming than did the experts. This latter finding is generally consistent with the technology adoption literature (Morris, Venkatesh, & Ackerman, 2005), which shows that younger individuals are somewhat more likely to be early adopters of new technologies.

Manipulations and Measures

Training Scenario Content and Distraction. The simulation scenarios focused on two medical conditions that commonly present to the pediatric Emergency Department: prolonged seizure (status epilepticus) and severe allergic reaction (anaphylactic shock). In the status epilepticus scenarios, the case featured an infant who presented with seizure. The participants were required to stabilize the patient by following an appropriate sequence of clinical care. This first involved managing the patient's airway, breathing, and circulation. Next, the participant needed to break the seizure by administering the correct medication. Finally, the patient stopped breathing which required further airway management and breathing tube placement. In the anaphylaxis scenarios, the patient presented with signs and symptoms of a severe allergic reaction. Again, the participant was required to stabilize the patient by following an appropriate sequence of clinical care. Briefly, this involves administering a "cocktail" of medications to treat the allergic reaction. However, the patient's condition worsened, thereby requiring a cricothyroidotomy (a rare but high stakes procedure that involves making an incision to open the airway), as well as blood pressure medication to stabilize

the patient's vital signs. The level of task difficulty was kept constant (high) for each scenario to provide all participants with an appropriate level of challenge (Chang et al., 2019). The anaphylaxis scenario had 9 critically required actions and the seizure scenario had 8 actions; failure to complete a required action on time ended the scenario as a failure or error. Task distraction was systematically manipulated by altering the volume, type, and amount of ambient noises such as device alarms, PA system announcements, and crying parents in the treatment bay.

Self-Reported Cognitive Load. Immediately after each scenario was complete, participants self-reported their level of workload using the NASA Task Load Index (TLX) (Hart & Staveland, 1988). The TLX is a questionnaire-based measure of workload that addresses six distinct components: mental demands, physical demands, temporal demands, performance, effort, and frustration. The participants responded using a 100-point scale that ranged from "Low" (0) to "High" (100). A total score was then generated for each scenario by summing across the six component scores. The mean correlation among the six TLX component scores was high (r = .70), which confirms that the components are not independent. We therefore computed the mean TLX score per participant, per scenario, which is a generally accepted method for calculating workload using the TLX scale (Hart, 2006).

Unobtrusive Measures of Cognitive Load. Real-time measures of individual CL were computed approximately every 1.25 seconds using a series of custom-developed workload models and classifiers (Beaubien et al., 2019; Durkee et al., 2013; Pappada et al., 2016). The workload values, which were scored a 100-point scale, were derived using a combination of electroencephalogram (EEG) and echocardiogram (ECG) data, and were de-noised using accelerometry (motion) data. During a typical 7.5-minute simulation scenario, approximately 360 streaming CL measurements were recorded per participant. Because each participant's CL varied over the evolving clinical situation (see Figure 2), there was a great deal of within-scenario variability. To deal with this extremely large corpus of data (nearly 17,000 unique CL measurements across the entire sample), we computed the arithmetic mean level of CL for each participant during each training scenario, which is consistent with prior work (Dan & Reiner, 2017).

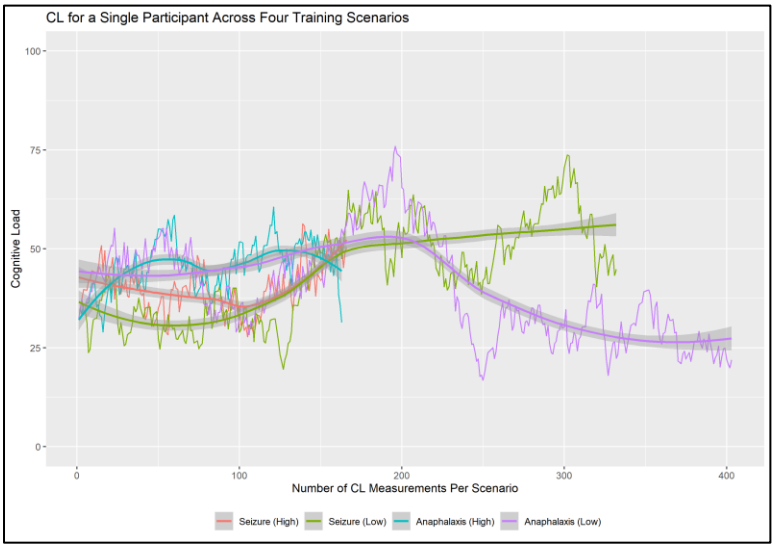

Figure 2. Graph of the streaming CL data for a single learner across all 4 simulation scenarios. A smoothing function has been superimposed to help visualize the trend in each scenario. In this example, the two high-distraction scenarios were completed much more quickly than their low-distraction counterparts.

Task Performance. The VR system recorded two main types of task performance measures: time-based measures and error counts. For the status epilepticus scenarios, the VR system recorded five time-based measures. These include the time to: suction the patient; apply oxygen; administer antiepileptic medications (both benzodiazepines and a longer-acting anti-seizure medicine); and intubate the patient. It also counted the number of task-related errors. Given how the clinical scenarios actually played out, complete performance data on all six measures were collected from every participant during every scenario. For the anaphylaxis scenarios, the VR system recorded six time-based measures. These include the time to administer certain medications (bronchodilators, antihistamines, epinephrine, pressors, and steroids), as well as the time to perform the cricothyroidotomy procedure. It also counted the number of

errors. Given how the actual clinical scenarios played out, there was a substantial amount of missing data for the blood pressure medications (pressors) and for the cricothyroidotomy procedure. This might occur, for example, if the patient crashed before the cricothyroidotomy procedure could be performed. When interpreting the task performance measures, it is important to note that the VR system contained several inherent limitations. As a result, some of the "errors" that were recorded by the system would not have been considered clinical errors in real life. For example, in real life the clinician would simultaneously perform multiple tasks and/or administer multiple medications. By comparison, the VR system artificially required that these tasks be performed sequentially and in a prescribed sequence. Any deviation from the simulator's prescribed sequence was treated as an "error."

RESULTS

Tutorial Times. The participants spent a mean of 178.57 seconds (SD = 24.87) using the tutorial to learn about the VR system's features and functionality before attempting their first clinical scenario. A two-way Analysis of Variance (ANOVA) revealed that there were no mean differences as a function of domain expertise ($F_{(1,11)}$ = 0.15, *ns*) or prior VR experience ($F_{(1,11)}$ = 0.90, *ns*). Therefore, tutorial time did not confound the results.

Streaming CL Data. To assess the effects of domain expertise and scenario content on streaming CL measures, we conducted a linear mixed model (LMM) on the mean streaming CL scores for each scenario. The results revealed a statistically significant main effect of expertise ($F_{(1,10)}$ = 7.20, *p*. <. 05) and a marginally significant expertise-by-scenario interaction ($F_{(5,50)}$ = 2.32, *p*. = .06). The fixed effects of expertise and scenario explained 29.8% of the criterion variance. Drill-down analyses suggest that experts had significantly lower mean CL levels than novices. The effect held for the baseline period, the tutorial scenario, and the anaphylactic shock scenarios. However, the expertise main effect did not hold for the seizure scenarios (see Figure 3).

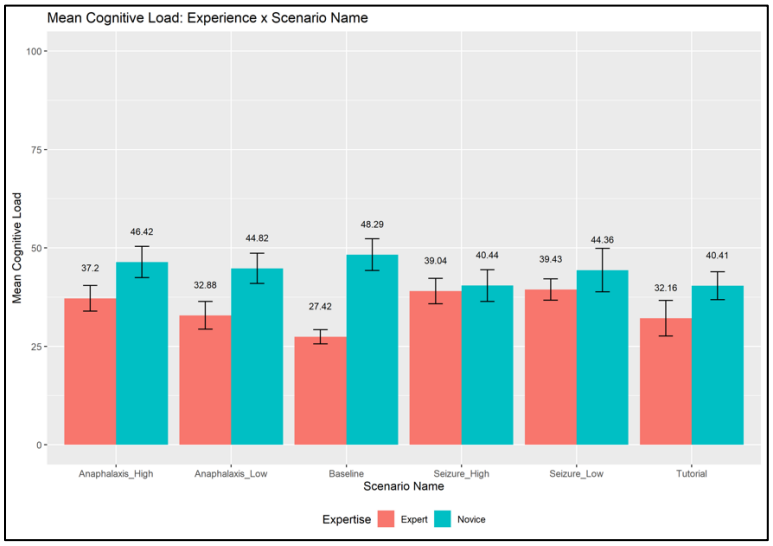

Figure 3. Graph of scenario mean streaming CL values. As depicted in the graph, experts had significantly lower mean CL values than novices. However, the effect did not hold for the seizure scenario.

Self-Reported Workload. To assess the effects of expertise and scenario content on self-reported workload, we conducted a linear mixed model (LMM) on the TLX total scores. The results revealed a statistically significant main effect of scenario ($F_{(4,40)}$ = 32.57, *p*. <. 01) and a significant expertise-by-scenario interaction ($F_{(4,50)}$ = 3.76, *p*. < .01). Unlike the streaming CL data, no main effect of expertise was detected [2]. The fixed effects model explained 45.8% of the criterion variance. With regard to the main effect of scenario content, drill-down analyses suggest this was caused by the tutorial scenario having the lowest workload values (see Figure 4).

[2] Similar effects were observed for the TLX mental demand component.

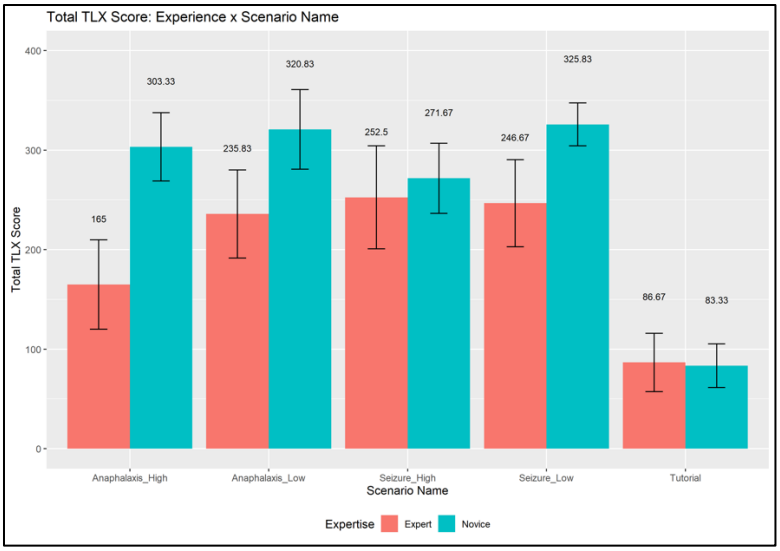

Figure 4. Graph of mean self-reported workload values. Many of the error bars overlap substantially. Of note, there were extremely high workload values for novices in the (low distraction) seizure scenario, and extremely low workload values for experts in the (high distraction) anaphylactic shock scenario.

With regard to the interaction effect, drill down analyses suggest that the effect was localized to the (high distraction) anaphylactic shock scenario. Curiously, despite the fact that many of the error bars overlapped considerably, in virtually every case the low distraction condition had higher means than the corresponding high distraction conditions. While this effect also held for the streaming CL data, that model did not detect a significant scenario main effect.

Task Performance. To assess the effects of expertise and scenario content on task performance, we conducted separate LMMs for the various completion time- and error-based performance measures. There were no statistically significant mean differences between domain experts and novices. This effect held for all scenarios and all performance measures. Because these null results were so counter-intuitive, we conducted a series of targeted drill-down analyses. For the first drill down analysis, we plotted the number of errors as a function of training scenario sequence to determine if there was evidence of learning (see Figure 5).

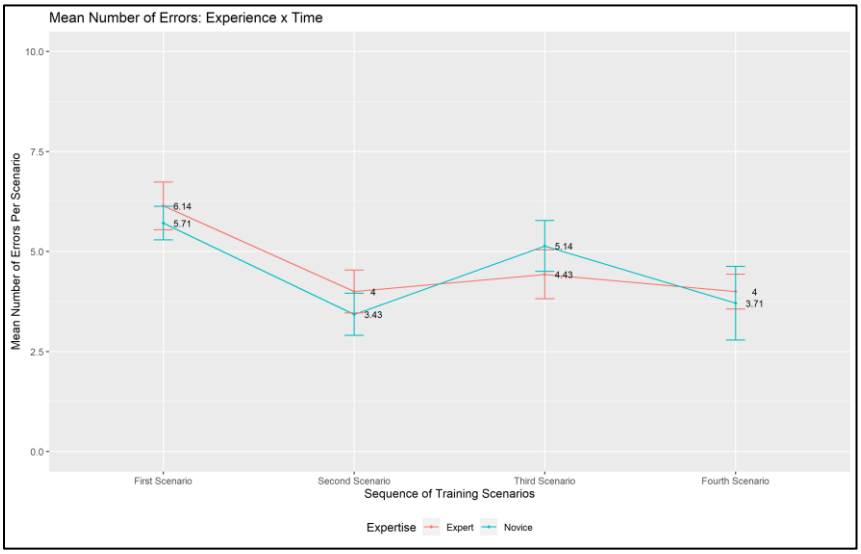

Figure 5. Graph of task-related errors as a function of simulation scenario sequence. For both expertise groups, the mean number of errors drops by nearly one-third across the four training scenarios. Because both groups improved over time, these results suggest that participants were learning to "play the simulator."

For both expertise groups, the number of errors decreased by nearly one-third over the four simulation scenarios. The results are confirmed by a LMM which identified a main effect of sequence ($F_{(3,48)} = 5.75$, $p. <. 01$), but no main effect of expertise, and no expertise by sequence interaction. Because both novices and experts improved over time, one must therefore infer that this reflects learning to "play the simulator," rather than learning the clinical tasks *per se*. If this was a true domain learning effect, the experts would have started with few errors and continued to exhibit few errors over time, while only the novices' number of errors would have decreased over time.

Similar visualizations and LMMs were conducted for all three measures of workload – the mean streaming CL measures (see Figure 6), the NASA-TLX mental demand scale (see Figure 7), and the NASA-TLX total score (see Figure 8) – to determine if there was evidence of learning.

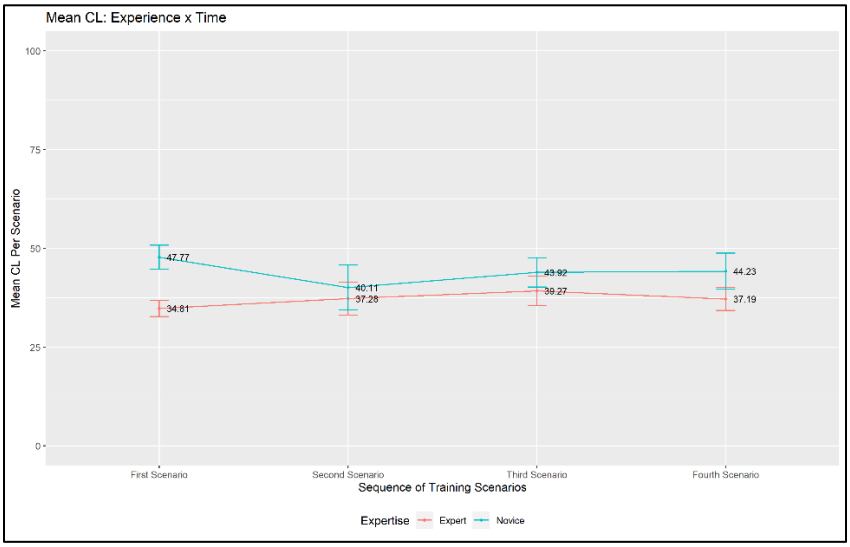

Figure 6. Graph depicting mean streaming CL values by expertise and scenario sequence. The results reveal a main effect of expertise, but no effect of sequence, and no expertise-by-sequence interaction.

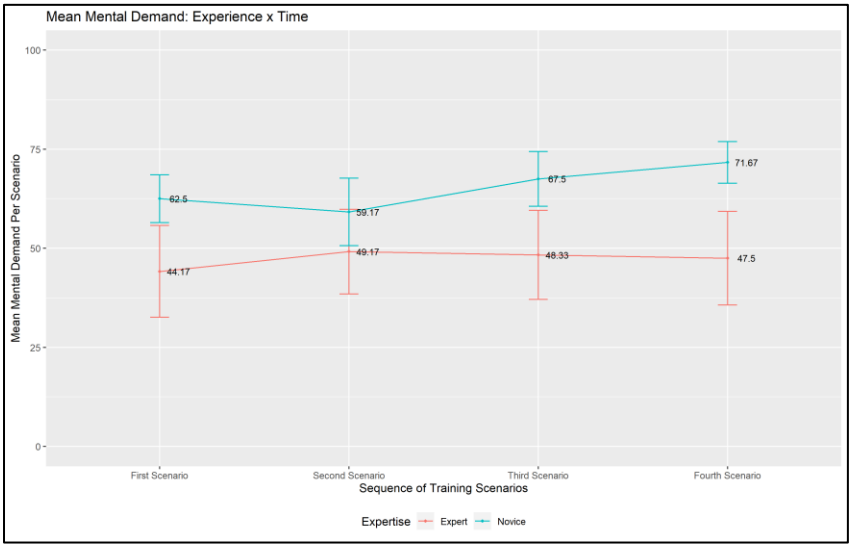

Figure 7. Graph depicting mean NASA-TLX mental demand scores by expertise and scenario sequence. The results reveal a main effect of expertise, but no effect of sequence, and no expertise-by-sequence interaction.

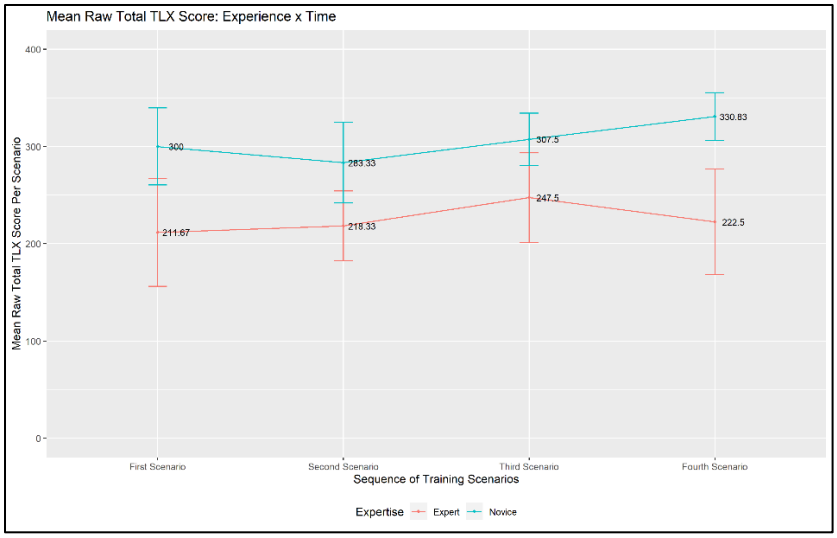

Figure 8. Graph depicting mean NASA-TLX total scores by expertise and scenario sequence. The results reveal a main effect of expertise, but no effect of sequence, and no expertise-by-sequence interaction.

In all three cases, the same pattern was observed. There was a statistically significant main effect of expertise, with experts having consistently lower levels of workload than novices. However, there was no significant main effect of scenario sequence, nor was there a statistically significant expertise-by-sequence interaction. In essence, the participants' workload did not decrease over time, suggesting that they were still using Type 2 decision processes at the end of the study. However, it is curious to note that the mean streaming CL values did have substantially smaller standard errors than either of the two self-reports. To the extent that this particular characteristic holds up in future research studies, it may permit more precise comparisons than self-reported methods.

CONCLUSIONS AND RECOMMENDATIONS

The results of this study were promising, but mixed. To recap, the research team leveraged real-time, unobtrusive measures of individual CL using a combination of EEG, ECG, and accelerometry data that were collected using wireless COTS neurophysiological monitors. We then conducted a small-scale validation study with known groups of novice and expert pediatricians who performed a series of realistic clinical simulations to compare the streaming CL workload measures with the NASA TLX scale. The two sets of workload data produced similar, but not identical, patterns of results. Specifically, when aggregating the results across all four scenarios, both metrics detected the expertise-by-scenario interaction, and both localized the expert-novice differences to the anaphylaxis scenario (see Figures 3-4). However, the streaming CL data detected the expertise main effect, while the TLX measures detected the scenario main effect. The exact causes behind these differences remain unclear.

By comparison, when analyzing the results by scenario sequence, all three metrics detected an expertise main effect, with experts having lower workload than novices (see Figures 6-8). This study also detected large, positive correlations among the TLX components (mean r = .70). This finding has been observed elsewhere (Flägel, Galler, Steinhäuser, & Götz, 2019; Lowndes et al., 2020; Tubbs-Cooley, Mara, Carle, & Gurses, 2018). Curiously, while the TLX components were highly intercorrelated, the mean streaming CL scores were not significantly correlated with the TLX mental demand component (r = .04, *ns*). However, the mean streaming CL scores were significantly correlated with the TLX task performance (r = .33, p. <. 01), effort (r = .27, p. <. 01), and frustration components (r = .27, p. <. 01), respectively. In the current study, these results make sense because the actual resuscitation task is fairly basic and should have been mastered by all participants, even the novices. In practice, the challenge is to perform the task in a noisy and distraction-filled environment. Taken together, the results suggest the need to look beyond simple Pearson correlations, and instead compare overall pattern of results when interpreting different operationalizations of the same construct, such as CL.

This study demonstrated that it is possible to measure CL unobtrusively using relatively low-cost COTS technology, and to make reasonable inferences about learner workload with some degree of certainty. In particular, the study

demonstrated that the streaming CL measure had smaller standard errors than the self-reported measures, which might prove useful in future studies since it may permit the detection of more fine-grained effects. In addition, future studies could leverage real-time CL alerts, for example to dynamically modify the training scenario content when a learner becomes significantly over- or under-loaded. Alternatively, the technology could be used – along with unobtrusive performance metrics – to help instructors baseline learners vis-à-vis their cohorts. Currently, medical residents are promoted to the next level based on their measured task performance, typically on an annual basis. However, the addition of CL values might be helpful in identifying those specific clinical tasks where residents could benefit from additional remediation even if their task performance is satisfactory, because it has not fully reached the point of automaticity. Other potential uses include the comparison of mean workload values when new technologies are inserted into the workplace, and/or new clinical practice guidelines or standard operating procedures are being developed. New technologies, policies, and procedures will invariably lead to short-term workload increases simply because they require un-learning the habitual ways of doing things. However, this effect should be temporary, and the heightened CL values should decrease over repeated trials. To the extent that CL values do not decrease over time, it might suggest that the new technology or approach is inferior to the old one, at least with regard to workload.

Finally, the VR simulator's time- and error-based performance measures failed to differentiate among the known groups of expert and novice physicians. While the number of errors did decrease over the four trials, they were still fairly high, given the relatively brief length (approximately 7-10 minutes) of each scenario. This finding, coupled with the fact that CL did not decrease over training scenarios, suggests several things. First, it suggests that the VR systems' performance measures do not reflect clinical errors *per se*, but rather they reflect errors as determined by the simulator's limited logic model. Second, it suggests that all participants were still operating using Type 2 decision processes even during the fourth and final clinical scenario. Third, and finally, it suggests that the simulator – perhaps caused by the system's user interface – was imposing an extraneous cognitive load (Van Merriënboer & Sweller, 2010) on learning to perform the task. Over repeated trials, the extraneous cognitive load may lessen. However, this time-limited study did not provide sufficient training trials to determine at what point that might occur.

LESSONS LEARNED

In the following paragraphs, we present several lessons learned that may benefit other researchers who are attempting to include streaming measures of individual CL in their own research efforts. Additional lessons learned can be found in Beaubien et al. (2019).

Lesson Learned #1: Measure CL and task performance simultaneously, because knowledge of one without the other can be very misleading. Imagine a theoretical 2 (CL) by 2 (task performance) matrix. Experts would demonstrate low levels of CL and high levels of task performance. Intermediate performers would demonstrate high levels of CL and high levels of task performance. Novice performers would demonstrate high levels of CL and low levels of task performance. Finally, participants who are not taking the simulation seriously would demonstrate low levels of CL and low levels of task performance. The critical point is that it is difficult to interpret either measure in isolation.

Lesson Learned #2: Recognize that new instructional media may impose an extraneous cognitive load on the learning process, but that this should lessen over time. All of the participants in this study demonstrated low levels of familiarity with VR and gaming technology. Moreover, all spent only a few minutes learning how to operate the VR system. In retrospect, it is not surprising that they all demonstrated fairly high levels of CL across the sequence of four training scenarios. In practice, it may require several practice trials for learners to master the user interface of a new training technology, so that they can devote their full attention to learning the clinical task.

Lesson Learned #3: Carefully consider the limitations of different types of human performance measures. The VR system generated a finite number of task performance measures, many of which were time-based. While evidence based-medicine has clearly demonstrated that certain time-based metrics (such as "door to balloon" time when the patient has blocked coronary arteries) are associated with positive clinical outcomes, time-based measures are less useful for training purposes than are process-based measures and process-based feedback. As a general rule, time-based metrics fail to inform the learner what they did right or wrong, thereby making it difficult to provide remedial feedback. Process-based metrics are much more useful when it comes to providing remedial feedback (Lefroy, Watling, Teunissen, & Brand, 2015; Shute, 2008).

Lesson Learned #4: When using multiple measures of the same construct, look beyond simple Pearson correlation coefficients. As demonstrated in the current study, the Pearson correlation coefficients among the mean streaming CL values and the TLX mental demand, physical demand, and temporal demand components (generated across the baseline scenario, the tutorial scenario, and the four clinical scenarios) were non-significant. However, the mean streaming CL values were significantly correlated with the TLX performance, effort, and frustration components. Moreover, the overall pattern of results – which detected the interaction effect for the performance-based analysis in Figures 3-4, and the expertise main effect for the learning-based analyses in Figures 6-8 – was very similar across the CL measurement methods. Therefore, researchers should carefully review the entire pattern of results, not just a single correlation coefficient.

Lesson Learned #5: Carefully consider how to operationalize concepts such as "task complexity" and "task distraction." For the purposes of this study, we experimentally manipulated task distraction rather than task complexity. Because we felt that it was necessary to maximize the VR simulator's task difficulty to maintain the learners' attention, the only things left to vary were the amount and type of extraneous environmental stimuli. While external distractions will invariably compete for the learner's limited attentional resources, they don't necessarily affect the task's complexity *per se*. Task complexity can be manipulated in three primary ways: 1) by increasing the number of distinct acts that must performed to complete the task, and/or the number of distinct information cues that must be attended to (component complexity); 2) by modifying the relationships among task cues, acts, and products (coordinative complexity), and; 3) by modifying the relationships among task inputs and products over time (dynamic complexity) (Wood, 1986). In retrospect, our task distraction manipulation was not a true manipulation of component complexity because it did not require the learners to attend to additional relevant decision cues; instead, it merely required them to ignore irrelevant environmental cues.

ACKNOWLEDGEMENTS

This study was supported with Internal Research and Development (IRAD) funds provided by Aptima, Inc.'s Office of Science and Technology (OS&T) and was approved by the Institutional Review Board (IRB) at Columbia University Vagelos College of Physicians and Surgeons. The authors acknowledge the VR simulations in this study which were developed by A.i.Solve, Ltd. (Luton, UK) and BioFlightVR (Culver City, USA) through the Oculus from FaceBook VR for Good Campaign. The authors would like to thank the pediatricians who volunteered to participate in our study, as well as the Department of Emergency Medicine at Columbia University for providing the research team with an Oculus Rift system.

REFERENCES

Beaubien, J., Wiggins, S., & DePriest, W. N. (2019). Real-time measurement of team cognitive load during simulation-based training. Paper No. 19129. In *Proceedings of the 2019 Interservice/Industry Training, Simulation, and Education Conference*. Alexandria, VA: National Training and Simulation Association.

Chang, T. P., Beshay, Y., Hollinger, T., & Sherman, J. M. (2019). Comparisons of stress physiology of providers in real-life resuscitations and Virtual Reality–simulated resuscitations. *Simulation in Healthcare, 14*(2), 104-112.

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

Dan, A., & Reiner, M. (2017). Real time EEG based measurements of cognitive load indicates mental states during learning. *Journal of Educational Data Mining, 9*, 31-44.

Dreyfus, S., & Dreyfus, H. (1980). *A five-stage model of the mental activities involved in directed skill acquisition* Technical Report No. ORC 80-2. Operations Research Center, University of California - Berkeley. Berkeley, CA.

Durkee, K., Geyer, A., Pappada, S., Ortiz, A., & Galster, S. (2013). *Real-time workload assessment as a foundation for human performance augmentation.* Paper presented at the International Conference on Augmented Cognition.

Evans, J. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*, 454-459.

Evans, J., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives in Psychological Science, 8*, 223-241.

Flägel, K., Galler, B., Steinhäuser, J., & Götz, K. (2019). The National Aeronautics and Space Administration-Task Load Index (NASA-TLX) - An instrument for measuring consultation workload within general practice: Evaluation of psychometric properties. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 147*, 90-96.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), 236-243.

Gray, R. (2017). Movement automaticity in sport. In J. Baker & D. Farrow (Eds.), *Routledge Handbook of Sport Expertise* (pp. 74-83). New York, NY: Routledge.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX): 20 years later. In *Proceedings of the 2006 Annual Meeting of the Human Factors and Ergonomics Society* (pp. 904-908). Thousand Oaks, CA: Sage Publications.

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

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*, 515-526.

Lefroy, J., Watling, C., Teunissen, P. W., & Brand, P. (2015). Guidelines: The do's, don'ts and don't knows of feedback for clinical education. *Perspectives on Medical Education, 4*(6), 284-299.

Lowndes, B. R., Forsyth, K. L., Blocker, R. C., Dean, P. G., Truty, M. J., Heller, S. F., . . . Nelson, H. (2020). NASA-TLX assessment of surgeon workload variation across specialties. *Annals of Surgery, 271*(4), 686-692.

Morris, M. G., Venkatesh, V., & Ackerman, P. L. (2005). Gender and age differences in employee decisions about new technology: An extension to the theory of planned behavior. *IEEE Transactions on Engineering Management, 52*(1), 69-84.

Pappada, S. M., Papadimos, T. J., Lipps, J. A., Feeney, J. J., Durkee, K. T., Galster, S. M., . . . Castellon-Larios, K. (2016). Establishing an instrumented training environment for simulation-based training of health care providers: An initial proof of concept. *International Journal of Academic Medicine, 2*(1), 32-40.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Tubbs-Cooley, H. L., Mara, C. A., Carle, A. C., & Gurses, A. P. (2018). The NASA Task Load Index as a measure of overall workload among neonatal, paediatric and adult intensive care nurses. *Intensive and Critical Care Nursing, 46*, 64-69.

Van Merriënboer, J. J., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. *Medical Education, 44*(1), 85-93.

Wood, R. E. (1986). Task complexity: Definition of a construct. *Organizational Behavior and Human Decision Processes, 37*, 60-82.

Wulf, G., McNevin, N., & Shea, C. H. (2001). The automaticity of complex motor skill learning as a function of attentional focus. *The Quarterly Journal of Experimental Psychology, 54A*(4), 1143-1154.