Real-Time Measurement of Team Cognitive Load During Simulation-Based Training

Jeffrey M. Beaubien, Ph.D.[1], Sterling L. Wiggins, M.A.[2], William Noah DePriest, B.S.[2]
Aptima, Inc.
Woburn, MA[1], Fairborn, OH[2]
jbeaubien@aptima.com, swiggins@aptima.com, wdepriest@aptima.com

ABSTRACT

The "Zone of Proximal Development" (ZPD) (Vygotsky, 1978) represents the difference between a learner's current and potential levels of mastery. While in the ZPD, learners experience facilitating levels of workload, motivation, and anxiety. During simulation-based training, instructors often attempt to keep learners in the ZPD by dynamically modifying the scenario's difficulty. However, accurately doing so requires real-time measures of individual and team Cognitive Load (CL), which have been heretofore unavailable. Using wireless commercial-off-the-shelf (COTS) neurophysiological monitors, we generated real-time measures of team Average Cognitive Load (ACL) and Cognitive Load Balance (CLB). We then integrated these measures with high-definition streaming video and expert observer ratings to provide a holistic view of team workload and performance. To test the veracity of our CL models, we conducted an initial validation study with 15 multidisciplinary medical teams. Three members of each team (the surgeon, the scrub nurse, and the anesthesiologist) were outfitted with the neurophysiological monitors. Each team then completed several clinical simulations of varying (low, medium, and high) difficulty. During a typical 12-minute scenario, we collected approximately 600 streaming ACL and CLB measurements per team. In addition, an expert observer rated the teams' technical proficiency and counted the number of appropriate and inappropriate clinical behaviors, respectively. Finally, at the end of each scenario, the team members self-reported their levels of anxiety and workload. The results suggest that as the teams' workload increased: their technical proficiency decreased; the number of appropriate clinical behaviors decreased; the number of inappropriate clinical behaviors increased; and their self-reported anxiety and workload increased. Although the magnitude of some effect sizes were small (Cohen & Cohen, 1983), all were in the hypothesized direction. Implications for future research and practical lessons learned are discussed.

ABOUT THE AUTHORS

Dr. Jeffrey M. Beaubien is a Distinguished Principal Scientist and Institutional Review Board (IRB) Chair at Aptima, Inc. 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 Research Technologies 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.

Mr. Sterling L. Wiggins is a Principal Scientist and Deputy Division Director in Aptima's Learning and Training Systems division. He leads several projects that develop technology and training solutions for operators in high-risk, safety-critical environments. His research interests include Live, Virtual, and Constructive (LVC) training, human-automation interaction, and adaptive aiding. He holds an M.A. in Education with a focus on learning, design, and technology from Stanford University, and a B.S. in Psychology from the Ohio State University.

Mr. William Noah DePriest is a Senior Software Engineer in Aptima's Performance Assessment and Augmentation Division. A large amount of Noah's work involves designing and developing software for human-system environments, including integration with physiological sensors as well as task simulators so that data can be collected, stored, processed, and visualized in both real-time and post hoc. He has also worked closely with scientists and

engineers to integrate advanced analytics and algorithms into software architectures to provide real-time assessments of operator state. He received his B.S. in Computer Engineering from Ohio Northern University.

Real-Time Measurement of Team Cognitive Load During Simulation-Based Training

Jeffrey M. Beaubien, Ph.D.[1], Sterling L. Wiggins, M.A.[2], William Noah DePriest, B.S.[2]
Aptima, Inc.
Woburn, MA[1], Fairborn, OH[2]
jbeaubien@aptima.com, swiggins@aptima.com, wdepriest@aptima.com

BACKGROUND

Since the publication of *To Err is Human* (Kohn, Corrigan, & Donaldson, 1999), healthcare provider organizations have overwhelmingly embraced simulation-based team training as a strategy for reducing medical errors and improving the quality of patient care. While much theoretical and empirical research has been written on the topic (Alonso et al., 2006; Burke, Salas, Wilson-Donnelly, & Priest, 2004; Weaver, Dy, & Rosen, 2014; Weaver et al., 2010), the process remains largely formulaic in practice. After completing some didactic instruction on the principles of effective communication, coordination, and decision-making, the team completes one or more realistic clinical simulations, typically using some form of realistic patient manikin. During the simulation, the team's performance is recorded using a high-definition video camera, and their performance is assessed using a standardized rating scale or checklist. Afterward, an instructor leads the team in a Socratic debriefing to diagnose their strengths and weaknesses, and to identify critical lessons learned. The video recordings and performance ratings support this process of team self-correction by highlighting "critical events" that are worthy of discussion during the debriefing. Ideally, this process of simulation followed by self-reflection is repeated until the team's performance has reached a pre-defined criteria of mastery. In practice, however, the number of training trials is often limited by schedule, instructor, and simulator availability.

Although there is compelling evidence to support its effectiveness (Hughes et al., 2016; Weaver et al., 2014), simulation-based team training is an extremely resource-intensive instructional method. At a minimum, three learners are required to participate in the clinical simulation, along with an instructor who manages the training event. Often, the instructor is supported by a simulator technician who dynamically modifies the simulator characteristics in response to the instructor's guidance, and by an expert observer who rates the team's performance so that the instructor doesn't spend all of his or her time "heads down" documenting the process. The scheduling of team training events can be a particular challenge, especially when the participation of clinical specialists (e.g., anesthesiologists, radiographers) is required, or when the ratio of training resources (e.g., simulated operating rooms, debriefing rooms) to teams is low. In some cases, the simulation may be conducted *in situ* (Miller, Riley, Davis, & Hansen, 2008). However, doing so renders the location temporarily unavailable for clinical care. Given the substantial costs and logistical constraints associated with simulation-based team training, many healthcare teams receive only a single, multi-hour block of dedicated team training per year.

Therefore, it is incumbent that healthcare organizations provide the best possible learning experience for each team. In the field of developmental psychology, there is a concept called the "Zone of Proximal Development" (ZPD) (Vygotsky, 1978), which represents the difference between a learner's current and potential levels of mastery. To ensure an optimal learning experience, instructors should attempt to keep the learners within the ZPD at all times. Too much challenge (the "Zone of Confusion") results in frustration, while too little challenge (the "Zone of Boredom") results in apathy. Ideally, the instructor does this by dynamically modifying the simulation scenario, thereby keeping the scenario's level of difficulty slightly above the team's skill level. The ability to dynamically modify training scenario content has been limited by the inability of instructors to accurately measure learner cognitive states, such as team Cognitive Load (CL), in real-time. Fortunately, recent technological advances, such as the development low cost commercial-off-the-shelf (COTS) wireless neurophysiological monitors, have now made this possible.

The construct of CL is rooted in the dual-process theory of decision-making, which postulates that there are two cognitive processes that operate largely, but not completely, in parallel (Evans, 2003; Evans & Stanovich, 2013). One of these processes, "Type 1," operates entirely at the unconscious level. It 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. It is much slower, places heavy demands on cognitive resources such as working memory, and bases decisions on explicit calculations and deliberation. It is akin to the slow and deliberate decision making approach used by novices, as well as the slow and deliberate approach used by experts when facing novel problems or situations for which their expertise has not fully prepared them (Kahneman & Klein, 2009).

Being a relatively new construct, there is little agreement on how to measure team CL, even though integrating neurophysiological signals from the individual team members has been identified a promising method (Bedwell, Salas, Funke, & Knott, 2014; Funke, Knott, Salas, Pavlas, & Strang, 2012). If the team's task requires a low level of team member interdependence – where each individual works in isolation, and where team performance is computed as the sum of the individual team members' performance (Saavedra, Earley, & Van Dyne, 1993) – then one could compute the arithmetic mean of each members' CL. By comparison, for highly interdependent team tasks – where one team member's output becomes another team member's input, and this process continues iteratively until the entire task has been completed (Saavedra et al., 1993) – it may not make sense to simply average the CL values of each individual team member. Instead, one might calculate the distribution of team members' CL scores, in order to assess the extent to which one or more team members has become significantly overloaded vis-à-vis the rest of the team.

With this in mind, the research team developed two novel measures of team CL. The source data for both measures are streaming neurophysiological data – electroencephalogram (EEG), electrocardiogram (ECG), and motion (accelerometry) – that are collected from the individual team members. The first measure is called "Average Cognitive Load" (ACL), which we define as the arithmetic mean of CL scores across an N-member team. It is particularly relevant when the task requires low team member interdependence. The second measure is called "Cognitive Load Balance" (CLB), which we define as the dispersion of CL scores across the N-member team. It is particularly relevant when the task requires high team member interdependence (Pappada et al., 2016). In both cases, a combination of EEG and ECG data were used to generate streaming team CL measures over an N-second window of time. The accelerometry data was used to "de-noise" the resulting workload scores for the effects of physical motion.

System Prototype

The research team developed a prototype hardware-software system that: integrates wireless electroencephalogram (EEG) and electrocardiogram (ECG) data streams from multiple team members; produces real-time measures of individual and team CL; uses accelerometer (motion) data to "de-noise" the participants' individual CL measures, and; graphically visualizes the results on a large-screen dashboard. The system is technology agnostic; it can ingest streaming data from any wireless COTS neurophysiological monitor. However, we have selected the BioRadio 150 sensor (Great Lakes Neurotechnologies, Cleveland, OH) because all three data streams (EEG, ECG, accelerometer) are sampled at the same frequency, thereby obviating the need to manually time-synchronize data streams that are recorded at different sampling rates. In the current system configuration, each participant wears a wet electrode EEG cap that is attached to the scalp and ECG electrodes that are applied to the chest (see Figure 1). Each BioRadio sensor streams the raw neurophysiological data to a separate Microsoft Surface Pro tablet for pre-processing.

Figure 1. A photo taken during early system testing. The wet electrode EEG cap is worn under the surgical bonnet. A flat ribbon wire transmits the EEG signal to the BioRadio sensor, which is attached to the participant's belt. ECG data is collected via electrodes that are attached to the chest.

The processed data – from all N team members – are then streamed to a PC server which calculates the ACL and CLB measures over an N-second window of time, visualizes the data on a large-screen dashboard, and provides real-time instructor alerts if the team is significantly over- or under-loaded vis-à-vis baseline (see Figure 2). The processed signals are also archived in a database for subsequent replay during a post-training debriefing. Two separate dashboards were developed: one for runtime visualization, and the other for the post-training debriefing. Both present real-time displays of team (upper left) and individual CL (upper right), both of which are depicted as partially-completed circles. When the CL values are within statistically normal bounds, the displays appear blue. When the CL values are statistically too low or high vis-à-vis the team's benchmark value, the displays appear red and flash to attract the instructor's attention.

Figure 2. The post-training debrief dashboard includes measures of team CL (upper left), high-definition video (upper middle), individual CL (upper right), and telemetry-style visualizations of user-selected CL metrics (center). The yellow vertical line can be used to "scroll" through the data, which are time-synchronized.

The bulk of both dashboards is devoted to presenting telemetry visualizations of the individual and team CL metrics. Three rows of data can be presented at a time, and filters can be used to determine what information to display. Directly above the telemetry data are a series of icons that correspond to expert observer ratings, which are collected via an Android tablet. These ratings provide contextual information that is otherwise absent from the individual and team CL measures, such as overall assessments to team skill/proficiency, or critical contextual cues (e.g., the patient crashes) to help interpret the team's performance during the debriefing. There are certain notable differences between the runtime and debriefing dashboards. For example, the runtime dashboard does not allow the user to "scroll" back through the data to specific points in time. It merely depicts what is currently happening, and provides a limited historical perspective of what has happened up until that point. By comparison, all of the data visualized on the debriefing dashboard – including the CL metrics, the high-definition video, and the expert observer ratings – are time synchronized. As a result, the instructor can "scroll" back and forth through the data and video to highlight critical events for the team to discuss.

PROCEDURE

The purpose of the current study was to validate the extent to which the team CL measures were working as intended. Specifically, we sought to answer the following three research questions: *To what extent do the ACL and CLB measures differentiate among scenarios of known "low," "medium," and "high" clinical difficulty?*, *To what extent do they correlate with external observer ratings of team performance?*, and *To what extent do they correlate with the team members' own self-reports?* Subsequent studies will assess the usefulness of the dashboard as a tool for helping instructors to dynamically modify training scenarios on-the-fly, and for supporting the post-training debriefings.

Method

After providing informed consent, 15 multidisciplinary medical teams each completed a series of surgical simulations. Three members of each team – the anesthesiologist, the surgeon, and the scrub nurse, all of whom operate within the sterile field – were fitted with a BioRadio sensor. Due to cost considerations, the other team members were not. The sensors were tested to ensure that they were properly communicating with the server, and that the individual and team CL metrics were accurately visualized on the runtime dashboard. Each team then performed a baseline scenario to benchmark their CL levels for comparison during the subsequent training scenarios.

The teams were then randomly assigned to complete multiple clinical scenarios (hemorrhage control, airway management, and trauma management) at varying levels of levels of technical difficulty (low, medium, and high difficulty). While complete counterbalancing was not possible due to logistical constraints, most teams (87%) completed between 4-6 training scenarios each, depending on their schedule availability. Usable data were collected from 71 training scenarios across the 15 teams. Each training scenario lasted approximately 10-15 minutes. Immediately afterward, the team members self-reported their level of workload using the NASA Task Load Index (TLX) (Hart & Staveland, 1988) and their level of anxiety using the State-Trait Anxiety Index (STAI) (Marteau & Bekker, 1992). The team members then participated in a post-training debrief which lasted approximately 10-15 minutes. During the debriefing, the simulation laboratory was reconfigured for the next training scenario. The process continued until the end of the scheduled data collection exercise, which lasted approximately 3 hours.

While they were performing the clinical scenarios, the teams' performance was digitally recorded using a high-definition (1080p) webcam. Additionally, an expert observer (an Emergency Medicine Physician) rated the team's performance using an Android tablet. The rating categories – physical skills, judgment, communication, medical knowledge, and professionalism – were drawn from a standard competency model that was developed by the Accreditation Council on Graduate Medical Education (ACGME) (Holmboe, Edgar, & Hamstra, 2016). The observer also counted the number of appropriate and inappropriate clinical behaviors that were performed by the team during each scenario[1]. All of the data – the streaming measures of individual and team CL, the high definition video feed, the expert observer ratings, the behavior counts, and the participant self-assessments – were digitally collected and stored in a database.

[1] Due to resource and scheduling constraints, expert observers were not available to rate every team on every clinical scenario.

Participants

Participant assignment to teams was quasi-random. It was based primarily on schedule availability, with the only stipulation that each team must include only one member from each clinical specialty: surgery, anesthesia, and nursing. The sample had a mean age of 35.25 (SD = 9.99) years; a mean of 6.32 (SD = 7.51) years of expertise in their area of specialization, and; a mean of 3.08 (SD = 2.60) years of experience participating in scenario-based training exercises. The surgeons and anesthesiologists were a relatively homogeneous group. All were residents, and there were roughly equal numbers of participants in the second, third, and fourth years of their respective residency programs. The scrub nurses had a mean of 13.06 (SD = 10.54) years of experience. There was a great deal of variation in the nurses' experience, as evidenced by the relatively large standard deviation-to-mean ratio. None of the teams had ever worked together as an intact unit, which is not surprising given that all the surgeons and anesthesiologists were residents. While it is not uncommon for advanced surgical or anesthesiology residents to work independently of their faculty members, teaching hospitals generally do not schedule surgical and anesthesiology residents together on the same surgical procedure.

Manipulations and Measures

Scenario Difficulty. Scenario difficulty was an experimentally-manipulated variable that consisted of three levels: low, medium, and high. The research team developed three different classes of clinical scenarios: hemorrhage control, airway management, and trauma management. Each scenario was then systematically modified to generate low-, medium-, and high-difficulty variants, thereby producing a library of nine custom-developed scenarios. For example, the baseline hemorrhage control scenario was made more difficult by increasing the rate of bleeding and making the source of the bleed difficult to locate. Similar approaches were taken to systematically modify the difficulty of the airway management and trauma management scenarios, respectively.

Team Cognitive Load. Streaming ACL and CLB measures were computed approximately every 1.25 seconds, thereby providing roughly 48 measurements per minute. During a typical 12.5-minute training scenario, approximately 600 measurements were recorded per measure, per team (see Figure 3).

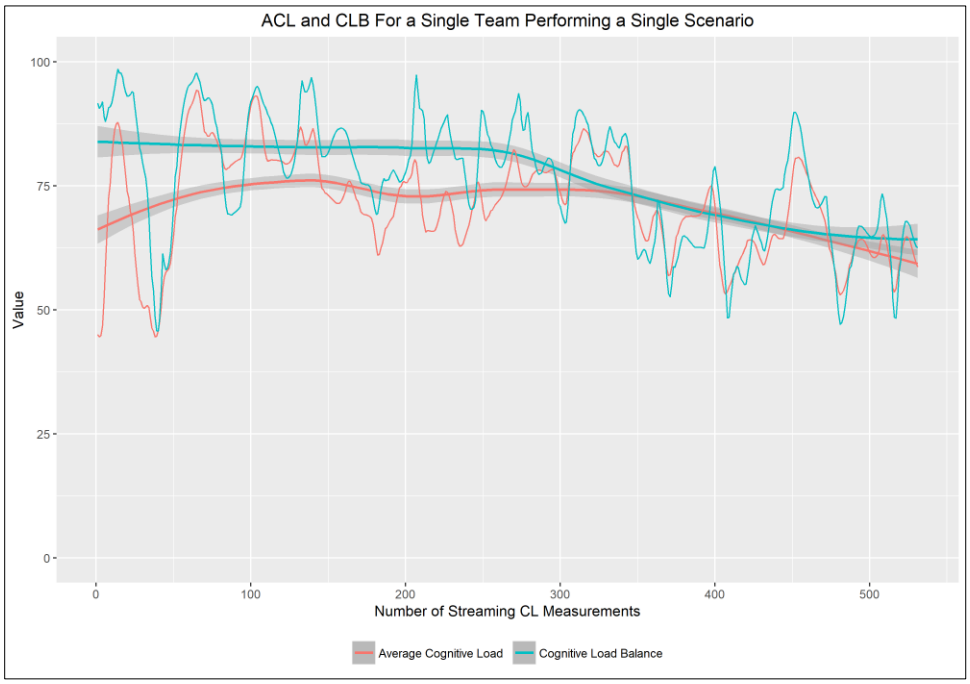

Figure 3. Graph of ACL and CLB data for a single team performing a single training scenario. Over 500 streaming measures of ACL and CLB were collected during this specific scenario. A smoothing function has been overlaid on the figure to help visualize this team's CL over the course of the scenario.

Both CL metrics were calculated on a 100-point scale (1 = low, 100 = high). High ACL values indicate that the team members are all relatively high in CL (the team was over-loaded). Conversely, low ACL values indicate that the team members are all relatively low in CL (the team was under-loaded). Moderate ACL values could result if the team members: had uniformly moderate levels of CL, or; if some team members had high CL while others had low CL. By comparison, high CLB values indicate that CL is balanced across the team members. For example, the team members' individual CL values could be uniformly high, uniformly moderate, or uniformly low. Low CLB values indicate that at least one team member was overloaded vis-à-vis the rest of the team. During a typical 12.5-minute training scenario, approximately 600 measurements were recorded per measure, per team. Given that each team's CL varied as a function of the evolving clinical situation, there was a great deal of within-scenario variability. To deal with this extremely large corpus of data (nearly 26,000 unique ACL and CLB measurements for the entire sample), we computed the arithmetic mean – separately for ACL and CLB – for each team during each scenario. Across the entire data set, per-scenario mean ACL and mean CLB scores were correlated $\underline{r} = .79$, indicating that the two measures were not independent.

ACGME Competency Ratings. While the teams were performing the simulation, an expert observer rated them on five of the six Accreditation Council on Graduate Medical Examination (ACGME) competencies (Holmboe et al., 2016): technical skills (patient care), judgment (patient care), medical knowledge, professionalism, and communication. Each competency was rated using a 100-point scale (1 = low, 100 = high), and multiple ratings were collected during each simulation. We computed the arithmetic mean of each competency for each team during each simulation event. A principal components analysis suggested the presence of a single component, which accounted for 97% of the total item variance. For the purpose of parsimony, we computed a single mean ACGME score per team, per scenario. This pattern of results – where expert ratings cluster by event, rather than by the specific skills being assessed – is consistent with prior research (Lance, 2008; Lance, Lambert, Gewin, Lievens, & Conway, 2004).

Behavioral Counts. The same expert observer also counted the number of appropriate and inappropriate clinical behaviors per team, per scenario. Both measures were positively skewed. Specifically, the count of appropriate clinical behaviors had a skewness value of 1.19. The count of inappropriate clinical behaviors had a skewness value of 3.70. As a general rule of thumb, skewness values greater than 1.0 are considered "highly skewed." Because attempts to normalize the data by converting to logarithms didn't appreciably improve the distributions, the variables were retained in their original scale. The correlation between the two counts was $\underline{r} = -.12$, n.s. As such, they were treated as separate dependent variables.

Self-Reported Workload. Immediately after each training scenario was complete, the three instrumented team members independently reported their level of workload using the NASA Task Load Index (TLX) (Hart & Staveland, 1988). The TLX is a self-reported measure of workload that addresses six conceptually distinct factors: mental demands, physical demands, temporal demands, performance, effort, and frustration. The TLX is scored in two steps. In the first step, each participant weights the relative importance of each factor based on its contribution to the total workload. In the second step, the participant rates the amount of each factor on a scale ranging from "Low" to "High." The weights from the first step are then multiplied by the values from the second step to derive a weighted total score, with larger values indicating higher levels of workload. A principal components analysis of the six TLX values suggested the presence of a single component which accounted for 64% of the total item variance. We therefore computed the arithmetic mean of weighted TLX scores (across all three team members) per team, per scenario.

Self-Reported Anxiety. The team members also independently reported their level of state anxiety using the State-Trait Anxiety Index (STAI) (Marteau & Bekker, 1992), immediately after each training scenario was complete. The STAI is a six-item, self-reported measure of state-based anxiety which inquires about both anxiety-present (tense, upset, worried) and anxiety-absent (calm, relaxed, content) emotions. After reverse coding the anxiety-absent items, the six questionnaire items are summed to generate a total score, with larger values indicating higher levels of anxiety. A principal components analysis of the six STAI items suggested the presence of a single component which accounted for 77% of the total item variance. We therefore computed the arithmetic mean of STAI items per team, per scenario.

RESULTS

Manipulation Checks

The TLX and STAI measures were moderately correlated ($r = .49$, $p. < .01$), which suggests that they should be treated as separate dependent variables. To ensure that the low-, medium-, and high-difficulty scenarios were working as intended, we separately computed one-way Analysis of Variance (ANOVAs) for the TLX and STAI measures, respectively. There was a statistically significant effect of scenario difficulty on self-reported workload, $F (2,68) = 5.52$, $p. < .01$. However, a subsequent Tukey HSD paired comparison test only revealed a significant difference between the low- vs. high-difficulty scenarios. The other paired comparisons (low- vs. medium-difficulty, and medium- vs. high-difficulty) were not significantly different from one another.

Similar results were observed for the STAI. There was a statistically significant effect of scenario difficulty on self-reported anxiety, $F (2,68) = 3.42$, $p. < .05$. Again, a subsequent Tukey HSD paired comparison test only revealed a significant difference between the low- vs. high-difficulty scenarios. The other paired comparisons (low- vs. medium-difficulty, and medium- vs. high-difficulty) were not significantly different from one another. Taken together, these results suggest that the scenario difficulty manipulation was only working at the extremes.

Scenario Difficulty

To assess the effect of scenario difficulty on team CL, we separately computed a one-way Analysis of Variance for the ACL and CLB measures, respectively. There was no statistically significant effect of scenario difficulty on ACL, $F (2,68) = 0.83$, n.s. There was also no statistically significant effect of scenario difficulty on CLB, $F (2,68) = 0.84$, n.s. These results suggest that the scenario difficulty manipulation did not produce systematic changes in mean ACL or CLB. However, given the results from our manipulation checks, these findings are not entirely unexpected.

Drill-down analyses suggest that the effects of scenario difficulty were not consistent across the three scenario content areas (see Table 1). Only the airway management scenarios exhibited clearly-defined and sequentially-ordered mean differences among the low-, medium-, and high-difficulty variants. By comparison, in the hemorrhage control scenarios, the means for the medium- and high-difficulty scenarios were virtually indistinguishable from one another. Finally, in the trauma management scenarios, the ordering of means between the low- and medium-difficulty variants were reversed. These results suggest that our scenario design and pilot testing efforts were not entirely successful.

Table 1. Means and Standard Deviations for Team Cognitive Load by Scenario Content and Difficulty Level

Scenario Content and Difficulty Level	ACL M (SD)	CLB M (SD)
Airway (Low)	59.41 (17.95)	68.36 (19.07)
Airway (Medium)	65.81 (20.70)	74.14 (18.26)
Airway (High)	72.03 (20.19)	78.29 (16.14)
Hemorrhage (Low)	64.82 (15.91)	66.55 (20.39)
Hemorrhage (Medium)	67.96 (16.73)	75.64 (14.75)
Hemorrhage (High)	68.71 (17.61)	74.96 (15.20)
Trauma (Low)	66.93 (20.22)	74.89 (19.92)
Trauma (Medium)	60.63 (18.30)	70.31 (16.46)
Trauma (High)	80.86 (11.93)	83.70 (12.89)

Expert Observer Ratings

To assess the effect of team CL on team performance, we separately correlated the per scenario mean values of ACL and CLB with the per scenario mean expert observer ratings. The results revealed a “large” effect (Cohen & Cohen,

1983) for both ACL (r = -.58, p. < .001) and CLB (r = -.47, p. 01), respectively. As the teams' workload increased, their technical proficiency decreased.

Expert Observer Behavior Counts

To further assess the effect of team CL on team performance, we separately correlated the per scenario mean values of ACL and CLB with the counts of appropriate and inappropriate clinical behaviors. The results revealed "small" effects (Cohen & Cohen, 1983), but ones that were in the hypothesized direction. For ACL, as workload increased, the number of appropriate clinical behaviors decreased (r = -.17, n.s.), and the number of inappropriate clinical behaviors increased (r = .23, p. < .10). For CLB, as workload increased, the number of appropriate clinical behaviors decreased (r = -.26, p. <. 05) and the number of inappropriate clinical behaviors increased (r = .20, n.s.). Given the heavily skewed distributions of these two dependent variables – most teams had small counts of each behavior type, but some teams had very large outliers – these findings are not unexpected.

Participant Self-Reports

To assess the effect of team CL on self-reported workload, we correlated the per scenario mean values of ACL and CLB with mean weighted TLX score per team, per scenario. The results revealed "small" effects (Cohen & Cohen, 1983), but ones that were in the hypothesized direction. For ACL, as workload increased, self-reported workload increased (r = .13, n.s.). Similar findings were observed for CLB. As workload increased, self-reported workload increased (r = .13, n.s.).

To assess the effect of team CL on self-reported anxiety, we correlated the per scenario mean values of ACL and CLB with the mean STAI score per team, per scenario. The results again revealed "small" effects (Cohen & Cohen, 1983), but ones that were in the hypothesized direction. For ACL, as workload increased, self-reported anxiety increased (r = .27, p. <. 05). Similar findings were observed for CLB. As workload increased, self-reported anxiety also increased (r = .15, n.s.).

CONCLUSIONS AND LIMITATIONS

The results of this initial validation study were promising, but mixed. To recap, the research team developed real-time, unobtrusive measures of team Average Cognitive Load (ACL) and Cognitive Load Balance (CLB) using a combination of EEG, ECG, and accelerometry data that were collected using wireless COTS neurophysiological monitors. We then conducted an initial validation study with multidisciplinary healthcare teams who were performing a series of realistic clinical simulations (hemorrhage control, airway management, and trauma management) at varying levels of clinical difficulty. The results suggest that as the teams' workload increased: their technical proficiency decreased (a large effect); the number of appropriate clinical behaviors decreased (a small effect); the number of inappropriate clinical behaviors increased (a small effect); and their self-reported anxiety and workload increased (a small effect). All of the effects were in the hypothesized directions.

The participant self-report measures (TLX and STAI) and expert observer ratings (ACGME ratings) all seemed to be working as intended. They demonstrated acceptable psychometric properties, such as having unidimensional factor structures. By comparison, the manipulation checks and CL analyses suggested that our scenario difficulty manipulations were not working entirely as intended. In particular, it appears that the medium-difficulty scenarios need to be revised such that they more clearly differentiate between the low- and high-difficulty scenarios, respectively. Part of the challenge may have been due to the nature of the participant sample. Because of logistical constraints, we were unable to recruit fellows and attending physicians. As a result, some of the findings may be caused by the fact that the teams – being largely composed of medical and surgical residents – were largely operating using Type 2 decision processes. In future research, we will attempt to expand our recruitment practices to obtain a more diverse sample of participants, including more expert clinicians, as well as more "expert teams" that have had multiple opportunities to work together in the recent past (Burke et al., 2004).

It is also possible that the very nature of streaming team CL measures requires fundamentally different data statistical analysis methods. As depicted in Figure 3, we often collected nearly 500 streaming team ACL and CLB values per team, per scenario. Moreover, there was a great deal of within-scenario variability as the team responded to various

challenges, such as when the patient "crashed" and had to be revived. Streaming CL data is not like a participant self-report, where there is a single score per team member, per scenario. Nor is it like an expert observer rating or a behavioral count, where there are a handful of observations per team, per scenario. Finally, it is not even like most simulator-derived measures – such as the number of enemies planes shot down during a single scenario. It's more akin to computing the mean airspeed of three jets during a 15-minute engagement, and then using this data to try and differentiate specific mission types –Air-to-Air vs. Close Air Support – from one another. Looking toward the future, we suspect that within-scenario comparisons may be more appropriate, such as comparing each team's streaming CL values during two clearly-defined "windows of time." For example, it may be insightful to focus on a clearly defined window of time that begins approximately 2-3 minutes before the patient unexpectedly crashes and ends 2-3 minutes after the team begins responding to the crisis. However, as noted in the introduction to this paper, team training presents a number of logistical challenges. To perform this kind of measurement, we would need to know the exact onset of the critical event; in this case, the exact moment the patient crashes. While we did eventually develop this recording capability using our Android tablet rating tool, we unfortunately could not recruit expert observers for every data collection. Moreover, those observers that we did recruit appeared to be task-saturated when making their other ratings, as evidenced by the high mean inter-correlation among the ACGME ratings. Fortunately, because all of the data – including the high-definition videos – have been archived, it may be possible to record the exact moment of each patient "crash" at a later date using an independent sample of clinical Subject Matter Experts.

LESSONS LEARNED

In the following paragraphs, we present several lessons learned that may benefit other researchers when attempting to compute streaming measures of individual or team CL in their own research efforts.

- *Guideline #1: Use a single COTS sensor to avoid having to time-synchronize multiple data streams that are recorded at different sampling rates.* Early during the system design process, the research team tested a number of potential COTS sensors. Eventually, we selected the BioRadio 150 because it includes multiple on-board sensors – including EEG, ECG, and accelerometry, among others – all of which are recorded at the same sampling rate. Had we chosen to use multiple different COTS sensors, our CL modeling efforts likely would have increased because of the need to time-synchronize the various data streams.
- *Guideline #2: Pilot test the scenario difficulty manipulations using the streaming CL measures along with participant self-reports.* We designed and developed nine custom simulation scenarios for this research effort, and pilot tested them by having candidate participants complete the NASA TLX after each one. Based on the pilot test results, we modified the scenarios accordingly to ensure adequate differentiation. However, development of the streaming CL measures was occurring simultaneously with the development of the clinical scenarios. As a result, we could only "eyeball" the real-time CL values for the different scenario difficulty manipulations. Looking toward the future, we intend to use both participant self-report and streaming CL measures when pilot-testing new scenarios.
- *Guideline #3: Given the large corpus of streaming data, consider replacing certain between-scenario statistical comparisons with within-scenario statistical comparisons.* As noted previously, there was a great deal of within-scenario variability as the team members responded to the evolving scenario conditions (see Figure 3). During future studies, we intend to compute team CL values at different "windows of time," such as immediately prior to and immediately following the onset of a critical event in the simulation. We believe that doing so will provide a clearer assessment of scenario difficulty levels. We have subsequently developed this functionality for our Android data collection tablet, which now allows the expert observer to identify the precise moment in time that the patient crashes using a single button press. We intend to use this functionality in future research efforts. Ideally, we would have integrated the CL measurement system with the simulator output. However, many patient manikins have "closed" system architectures that do not stream the data to external devices.
- *Guideline #4. With large enough teams, consider calculating separate CL models for specific sub-teams.* Due to resource constraints, we were only able to instrument three team members. However, in practice, there are multiple sub-teams in surgery: the anesthesia team (consisting of the anesthesiologist and nurse anesthetist), the surgical team (the surgeon, the surgical assistant, and the scrub nurse), and the support team (the circulating nurse, and the various technicians such as radiographers). In practice, these teams perform during different time cycles. For example, the anesthesia team first induces the patient. During this time, their workload is relatively high, but once the patient is fully sedated, their workload drops substantially. During

the induction phase, the surgical team plans the procedure. Once the patient is fully sedated, they begin the surgical procedure proper, during which time their workload increases substantially. The process then reverses itself as the surgical team completes the surgical procedure, and the anesthesiology team revives the patient. Surgery is a classic Multi-Team System (MTS), in which sub-teams pursue different proximal goals (often with different time tables) but have the same terminal goal (Klein & Kozlowski, 2000). Looking toward the future, we will attempt to acquire additional neurophysiological monitors, so that we can measure CL separately for each sub-team.

- *Guideline #5. If you want to provide real-time instructor alerts, the EEG signals needs to be pre-processed before streaming it to the dashboard.* One of our project goals was to provide real-time alerts to an instructor so that they could quickly determine if the team was in the "Zone of Confusion" or the "Zone of Boredom," and take appropriate action. In order to do so, we used Microsoft Surface Pro tablets to pre-process the streamed data from the BioRadio – to filter out artifacts and to derive features Heart Rate (HR), inter-beat intervals (IBI) and EEG frequency band powers – before pushing it to the server for run-time visualization. Without this pre-processing, we would not have been able to compute real-time models of individual and team CL.
- *Guideline #6: Operationally test your system in the intended site of deployment.* We collected our validation data set at two different academic teaching hospitals. While both data collections took place in dedicated simulation centers, de-conflicting multiple Bluetooth signals can be a challenge, especially in "noisy" environments. As a general rule, dedicated simulation centers tend to have less wireless and Bluetooth traffic than a clinical ward. As a result, if you intend to use your system in a dedicated clinical ward, be sure to pilot test it there. Don't pilot test the system in a relatively "clean" environment such as a simulation center, and then expect it to work *in situ* without modification.
- *Guideline #7. Be cautious when attempting to developing participant-specific CL "profiles."* A long-term goal of this effort is to develop personalized CL models for each clinician. We believe that doing so can potentially generate more accurate and diagnostic measures of individual and team CL. However, doing so will likely require developing a data warehouse with an extensive corpus of training trials (per participant) in order to identify the unique idiosyncratic characteristics of each clinician. Moreover, such personalized CL models will likely be unstable for anyone except extreme experts who are operating exclusively using Type 1 decision processes on well-defined tasks. Personalization would likely not work for novices, whose CL is evolving from Type 2 to Type 1 decision processes with increases in their domain expertise. Even for the experts, so-called "personalized" CL models would likely vary as a function of several factors, such as the clinician's physical state, whether the clinician is performing a routine vs. a new procedure, or whether the clinician is performing the task after an extended period of absence such as after taking extended family leave. The key point here is that while the development of personalized CL profiles is possible, it presents numerous practical challenges and will require extremely large volumes of data.

ACKNOWLEDGEMENTS

This work was supported by the US Army Medical Research and Materiel Command (USAMRMC) under Contract No. W81XWH-14-C-0021 to Aptima, Inc. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

The authors would like to thank the following individuals for their assistance during the process of system design, scenario development, and data collection: Dr. Scott Pappada (University of Toledo), Dr. Thomas Papadimos (University of Toledo), Dr. Susan Moffatt-Bruce (Ohio State University), and Mr. Jonathan Lipps (Ohio State University), as well as the numerous medical residents, anesthesia residents, and nurses who participated in our validation studies.

REFERENCES

Alonso, A., Baker, D., Holtzman, A. K., Day, R., King, H., Toomey, L., & Salas, E. (2006). Reducing medical error in the military health system: How can team training help? *Human Resource Management Review, 16*(3), 396-4154.

Bedwell, W. L., Salas, E., Funke, G. J., & Knott, B. A. (2014). Team workload: A multilevel perspective. *Organizational Psychology Review, 4*(2), 99-123.

Burke, C., Salas, E., Wilson-Donnelly, K., & Priest, H. (2004). How to turn a team of experts into an expert medical team: Guidance from the aviation and military communities. *BMJ Quality & Safety, 13*(Supplement 1), i96-i104.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* (2nd ed.). Mahwah, NJ: Erlbaum.

Evans, J. S. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. S. B. T., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives in Psychological Science, 8*(3), 223-241.

Funke, G. J., Knott, B. A., Salas, E., Pavlas, D., & Strang, A. J. (2012). Conceptualization and measurement of team workload: A critical need. *Human Factors, 54*(1), 36-51.

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.

Holmboe, E. S., Edgar, L., & Hamstra, S. (2016). *The milestones guidebook*. Chicago, IL: American Council on Graduate Medical Education.

Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., . . . Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology, 101*(9), 1266.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515-526.

Klein, K. J., & Kozlowski, S. W. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco, CA: Jossey-Bass.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (1999). *To err is human*. Washington, DC: National Academy Press.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology, 1*(1), 84-97.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*(2), 377.

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology, 31*(3), 301-306.

Miller, K. K., Riley, W., Davis, S., & Hansen, H. E. (2008). In situ simulation: A method of experiential learning to promote safety and team behavior. *The Journal of Perinatal & Neonatal Nursing, 22*(2), 105-113.

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.

Saavedra, R., Earley, P. C., & Van Dyne, L. (1993). Complex interdependence in task-performing groups. *Journal of Applied Psychology, 78*(1), 61.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34-41.

Weaver, S. J., Dy, S. M., & Rosen, M. A. (2014). Team-training in healthcare: A narrative synthesis of the literature. *BMJ Quality & Safety, 23*(5), 359-372.

Weaver, S. J., Salas, E., Lyons, R., Lazzara, E. H., Rosen, M. A., DiazGranados, D., . . . King, H. (2010). Simulation-based team training at the sharp end: A qualitative study of simulation-based team training design, implementation, and evaluation in healthcare. *Journal of Emergencies, Trauma and Shock, 3*(4), 369.