# Virtual Reality Provides Real Data: How Data in VR Transforms the Concept of Readiness

**Summer Rebensky, William Stalker, Shawn Turk, Samantha Perry**
**Aptima, Inc.**
**Fairborn, OH**
**srebensky@aptima.com, sturk@aptima.com, lstalker@aptima.com, sperry@aptima.com**

**Jonathan Diemunsch, Quintin Oliver, Winston "Wink" Bennett**
**Air Force Research Laboratory**
**WPAFB, OH**
**jonathan.diemunsch.1@us.af.mil, quintin.oliver@us.af.mil, winston.bennett@us.af.mil**

## ABSTRACT

NATO highlighted the possibilities of artificial intelligence (AI) for military applications: "The main ingredient in any ML-application is data…military organizations may have to adapt their data collection processes to take full advantage of modern AI-techniques" (NATO, 2018). To do so, we need to change the way we currently classify learning from multiple choice questions to objective data (Schatz & Walcutt, 2022). Before 2019, virtual reality (VR) in education relied on pre and post-tests, questionnaires, interviews, scales, focus groups, and observations (Çanakaya, 2019). These methods are labor intensive, consist primarily of qualitative approaches, and do not support the rich data needed to build AI into the training pipeline. In the past five years, researchers and engineers have been finding ways to incorporate objective data into VR systems. Human state examples include: eye tracking (Ahuja, et al., 2018), heart rate and respiratory rate (Floris et al., 2020), electroencephalogram (EEG) data (Tremmel et al., 2019), and facial expression tracking (Houshmand, 2020). Commercial off-the-shelf headsets have begun integrating human state data as well as cameras and sensors that provide opportunities to understand motion, where trainees spend time, their interactions, and every nuance of how they complete their training tasks. Tracking data within a VR training exercise provides insight into who needs additional training, when AI virtual instructors could assist, the level of readiness, the likelihood of trainee success, and detecting points of failure in the field to gaps in training. This paper will address the ambiguity of what kind of data to use for different training applications, how to pull data from VR systems, how the training pipeline could use this data, and examples of how data was built into various VR training tasks developed by an Air Force Research Laboratory with lessons learned and guidance.

## ABOUT THE AUTHORS

**Dr. Summer Rebensky**, Aptima Inc., is a Scientist in the Gaming Research Integration for Learning Laboratory (GRILL®). She has background focusing on human performance, cognition, and training in emerging systems working with Air Force, Navy, and FAA on training design, drone operations, aviation human factors, and gamified training. Her research leverages VR, AR, and technology to optimize human performance in training, Dr. Rebensky received her BA in psychology, MS in aviation human factors, and PhD in aviation sciences with a focus in human factors from Florida Tech.

**William (Liam) Stalker M.S.,** Aptima, Inc., is a Scientist Intern in the GRILL®. He uses his proficiency in neuroergonomics, human factors, and experimental psychology to the aid the Training, Learning, and Readiness Division. His efforts focus on adaptive training data collection efforts as well as survey methodology and implementation.

**Jonathan Diemunsch,** AFRL, is a Computer Scientist in the GRILL® with a decade of experience developing game tech for the Air Force to create training solutions. He worked to create a DIS plugin for better Air Force training. He attends tech conferences annually to stay on top and has contributed to national security efforts.

**Shawn Turk,** Aptima Inc., is an Associate Software Engineer in the GRILL®. and leverages a BS in digital simulation and game engineering technology from Shawnee State University to explore training solutions utilizing artificial reality and virtual reality technology. He has utilized Unreal Engine and Unity to incorporate sensor data to augment and adapt virtual tasking.

**Dr. Samantha (Baard) Perry**, Aptima, Inc. Senior Scientist in the GRILL® and has more than 10 years of academic and applied research experience with the Air Force, Army, NASA, and emergency medical teams, with expertise in adaptation, motivation, training design and evaluation, survey development and implementation, and unobtrusive measurement of team processes, states, and performance. Dr. Perry holds a PhD and MA in industrial and organizational psychology from Michigan State University and a BA in psychology from George Mason University.

**Quintin Oliver,** AFRL, is a Computer Scientist in the GRILL®. His work utilizes virtual, augmented, and mixed reality technologies to create rapid prototypes of environments focused on personalized training. In these environments, he leverages his interests of 3D modeling and artificial intelligence to create unique experiences.

**Dr. Winston "Wink" Bennett** is a Senior Principal Research Psychologist and Readiness Product Line Lead for the Warfighter Readiness Research Division, Airman Systems Directorate, 711th Human Performance Wing, Air Force Research Laboratory. Wink is a recognized leader in education, training, competency definition and assessment, and performance measurement research. He has been involved in a number of multinational research collaborations and continues to support collaborations around the world. He is a Fellow of three distinguished professional societies and the Air Force Research Laboratory.

# Virtual Reality Provides Real Data: How Data in VR Transforms the Concept of Readiness

**Summer Rebensky, Shawn Turk, William Stalker, Samantha Perry**
**Aptima, Inc.**
**Fairborn, OH**
**srebensky@aptima.com, sturk@aptima.com, lstalker@aptima.com, sperry@aptima.com**

**Jonathan Diemunsch, Quintin Oliver, Winston "Wink" Bennett**
**Air Force Research Laboratory**
**WPAFB, OH**
**jonathan.diemunsch.1@us.af.mil, quintin.oliver@us.af.mil, winston.bennett@us.af.mil**

## INTRODUCTION

Maintaining proficiency and readiness is critical for meeting Air Force mission requirements, and yet the process by which airmen and warfighters are required to show said readiness levels are time and resource heavy. Instructors need data about their trainees before they can properly assess and aid trainees. In traditional settings, instructors capture this data through pure observation or assessment. These methods often introduce biases with different instructors rating skills differently without standardization across multiple training programs. These methods can also prove to require a large amount of manpower and can detract from available training time or instructor time. In a data-centric era and with the growth of artificial intelligence (AI) capabilities, we have the potential to provide individualized, effective, training at all levels and environments. To do so, we need to change the way we currently classify learning from multiple choice questions to quality objective data (Schatz & Walcutt, 2022). The more finite the data, the greater potential to be able to classify, predict, identify, and address gaps in training. Before 2019, VR in education relied on pre and post-tests, questionnaires, interviews, scales, focus groups, and observations (Çanakaya, 2019). In 2023, there are multiple means of evaluating a learner's progress in the modern training environment before needing to resort to highly resource intensive and manual processes of pass/fail ratings or asking trainee's questions via assessments. Considering the frequent change in instructors and the shortened training timelines, providing instructors with rich (and automatically collected) data to support their role is key to optimizing the limited time available at training schoolhouses.

Researchers and instructors have explored using non-intrusive measures to assess a trainee's progress as either a supplement or potential substitute to the more traditional pen and paper evaluation process. An advantage of the non-invasive approach is that it allows instructors to seamlessly collect data from the developing operators, in an objective and hands-free way and even automatically adjust the delivery or difficulty of the curriculum in real-time; capturing trainee performance does not need to disengage from the flow of learning. Virtual reality (VR), and the data it provides, can enable quick views of the competencies, proficiencies, and readiness of a training unit as a whole. In simulation-based training, capabilities include: measuring elements related to each knowledge, skill, ability, and other characteristics (KSAOs), provide feedback in the moment, and collect many different skill metrics all tied back to learning objectives (Rebensky, Perry, & Bennett, 2022). Although traditional performance measurement has many logistical, resource, and collection constraints, these limitations are becoming smaller and smaller with each iteration of simulation and VR capability. Alternatively, if simulation-based training is used, computer programs can quantify the frequency of these interactions and provide a summary report at the completion of the task. Not all differences matter though. Data about the operator's behavior should only be analyzed if it pertains to the end outcome or goal of the task, or if it is related to the processes involved in achieving the goal of the task (Salas et al., 2005). Adding conceptually irrelevant data will muddle the performance analysis. Within this paper we present measures that have demonstrated relevance and feasibility within the training domain. We discuss the benefits of leveraging this data and how it can be used to inform training as well as examples developed in real-world training and research applications by an Air Force Research Laboratory.

## MODERN TRAINING DATA REQUIREMENTS

"Twenty minutes of VR use can generate approximately two million data points" (Bailenson, 2018, as cited in Jerome & Greenberg, 2021). So, what do we do with all that data? Considering the current state of AI, in 2023, requires

massive amounts of training data to be as informative as it is, that data can be gold for the training environment. Training AI like Chat-GPT3 and Chat-GPT4 required billions of text documents. To achieve the vision of readiness the modern warfighter requires, we will need similar quantities of information on trainees across the pipeline. The integration of AI models with learning management systems (LMS) brings transformative value to modern training programs. AI surpasses its human counterparts in its ability to handle massive amounts of data and perform advanced data analysis. It effortlessly processes diverse data formats, including structured data, unstructured text, images, and videos, without the need for extensive pre-processing. With the ability for AI models to connect to an LMS, it can seamlessly import and export data. This connection allows for the population of learner profiles and datasets, enabling personalized training recommendations based on individual learner profiles. It optimizes the learning experience, empowering training programs to adapt and align with the evolving needs of airmen. Moreover, the integration enables a dynamic feedback loop, continuously enhancing the AI model's accuracy and relevance through learner interactions and outcomes. This data-driven approach fosters efficient and personalized learning journeys, revolutionizing the training landscape.

The utilization of AI models in training will transform how training content is recommended and delivered. AI's seamless import and export capabilities, coupled with its connection to LMS, create a continuous flow of information between the AI model and the learning ecosystem. This integration will ensure the accurate capture of learner progress, preferences, and performance metrics, enabling personalized training recommendations. Department of defense (DoD) can leverage AI models to generate enterprise-level data-driven insights, populate learner profiles, and provide tailored training content. The iterative feedback loop between learners and AI algorithms drive continual improvement, adaptation and, optimizing the learning experience. These advancements in training data requirements will pave the way for efficient, adaptive, and personalized training programs that effectively align with the evolving needs of learners.

## RICH DATA SOURCES IN SIMULATION & TRAINING

Current data collection methods rely on instructor observations, surface level performance metrics, and pass-fail structures. Continuation training (CT), typically in the form of a generalized scenario, is the primary means for providing pilot training events necessary for maintaining proficiency (DAFMAN 11-401) but rarely show the variability of knowledge and skill capabilities of individuals, nor their acumen within the scenario. Considering the sensitivity needed to inform the readiness pipeline, passing grades at the schoolhouse level will not allow us to link back deficiencies observed in operations (e.g., why is Trainee A struggling during on-the-job training if both Trainee A and Trainee B got a “pass” grade in the block that covered this task? Likely one passed with soaring colors and the other barely passed, but retrospectively, we need the data to be sensitive enough to draw these conclusions). Non-intrusive measures broadly encompass means of assessment that neither requires the trainee to stop what they are doing, nor does it require them to report their thoughts and priorities during the training scenario. Non-intrusive measures instead capitalize on the information about the trainees’ physiological characteristics and/or their real-time performance in the training task. This data can then be used to better predict performance and improve training outcomes, develop AI that can match trainees with the ideal training content, beyond what is possible when relying only on basic instruction. For VR, sources of information can include: cameras, motion, physical environments, voice, biometric information, location, interactions, and usage (Jerome & Greenberg, 2021). These measures can be divided into three categories: the observation of their interactions with the system, the assessment of their performance within and outside of a team setting, and the monitoring of the trainee’s physiological state.

### Interactions with the System

Human factors specialists and researchers from similar fields will assess how an operator interacts with different components within the task environment. Some of the interactions they might assess in a driving task for example include: the frequency of speeding, frequency of aggressive acceleration followed by abrupt breaking, the sharpness of turns, and other similar measures. Each operator may complete the hypothetical goal of going from Point A to Point B, but assessing the subcomponent interactions within the system will reveal different levels of skills that pass / fail styled assessments will miss. These interactions can be quantified retroactively by reviewing recorded footage of the operator completing the task. Analyzing an operator’s system interactions provides insight that broad measures of performance frequently neglect. Using measures such as “number of hours flown,” may work as an indirect measure of a pilot's skill (Fletcher, 1999), but it fails to encapsulate more nuanced skills that make-or-break performance in

strenuous scenarios. To find these measures more entwined with performance outcome, instructors and researchers must ensure that they have adequate computational resources to detect and quantify all system interactions before implementation.

Operators are often performing a countless number of interactions during complex tasks that are too difficult for any human observer to quantify. This problem becomes exponentially harder as the subtlety of these interactions increases. One of the means of overcoming this dilemma is to pre-program the simulator to assess the meaningful interactions and provide generalized summaries about the trainee's performance for the instructor to review. Simulators must not only collect raw data about an operator's performance, they must also help make the data sensible to the end user. Instructors are substantially better at assessing an operator's training performance when simulators provide a meaningful aggregated interpretation of the system interactions (Ryder et al., 2003). Researchers collaborating with the Air Force Research Laboratory (AFRL) did just that when designing a flight task testbed called AGENT, the Agent Generation & Evaluation Networked Testbed (Freeman et al., 2019). This simulator was designed to collect information about the operator's interactions across many samples. This provides researchers and instructors with a rich data source that helps them identify the key associations between system interactions, training scenario modifiers, and the operator's performance. Designing simulators with these capabilities baked into the program reduces the burden on instructors and improves training efficiency.

State data within VR systems can provide granular information previously captured through observational means. Consider a maintenance task. An instructor may observe a trainee complete the replacement of a part within an aircraft. The instructors scoring sheet may include (from least to most granular): (a) a checklist that the instructor marks as each step is completed, (b) Qualifying data such as time to complete the task or some other performance notes, (c) in general the number of errors and whether or not assistance was needed—in some cases this could also be tracked per step, and (d) to what level of proficiency each step and safety standards were followed. At the higher levels, checklists can suffer from a lack of sensitivity, those that can successfully complete the tasks move on to future positions. If everyone gets 100%, how can we identify the potential deficiencies in either trainee readiness, or the training pipeline? Collecting more granular data through observational means requires standardized processes, manpower, and a method for entering the data within a centralized location. However, leveraging state data within the VR environment allows us to understand more granular information than ever. Positional information can provide great insight into how trainees explore environments. We can understand which parts of a learning environment that trainees spent the most time, where trainees may have gotten stuck, which elements within the environment that were within their field of view, and which areas of the environment were attended to. Interaction data can provide the proficiency-level data collected unobtrusively and objectively across trainees (i.e., specifically how good someone is at a specific task). Every object the trainee touches, picks up, interacts with, and actions they complete can be logged in the background automatically. Although this is also possible within some desktop-based simulations, VR can often leverage more naturalistic interactions that are more representative of the real-world task. We can measure the actual time it requires to set up a defibrillator by measuring the time to complete each step using the controllers as they would their hands, whereas with a desktop-based simulator or paper test we can only measure procedural-related knowledge. This level of granularity allows us to answer questions not only related to time to complete tasks and errors, but even reaction times, attempts, and smoothness of different interactions. Headset positional information can provide great abilities to collect kinematic data on the trainee (Spitzley & Karduna, 2019), whereas the controls can even provide extremely sensitive proficiency data information such as smoothness of maneuvers while controlling an aircraft (Emerson et al., 2022).

**Instructor, Observer, and Team Measures**

Assessing the performance of teams has been a staple in human factors research since the earliest developments in the field. Teams training research has taken place in variety of fields, such as aviation, healthcare, and military operations. A team's performance can be divided into outcomes and processes (Rosen et al., 2010). The outcome is the team's ability to achieve the goal. The processes are the behaviors executed to achieve said goal. An operator within a team can be assessed on the processes they contribute to but not necessarily the end outcome. Non-intrusive measures of an operator's performance in a team may consist of behaviorally anchored ratings by an observer (e.g., Lie et al., 2015) or performance in event-based assessments (e.g., Seelandt et al., 2014). These empirically backed techniques developed to help an instructors observe performance within a team context can also be used to provide individual feedback outside of a team setting. Behavioral anchored ratings scales are popular because of their ease. Rather than trying to discern the differences between a rating of a "2" versus a "3," these types of scales include descriptions for

each value on the continuum of the scale. For example, in context assessment of a presentation speaking skills, a score of a "1" might be reserved for "does not face the audience and reads strictly off of the presentation slides," while a "2" would be "faces the audience but does not elaborate beyond what is listed on the slides." In a VR environment, we can leverage the positions within the virtual environments, the chat functionality with natural language processing and the interactions they have with one another in the virtual environment. These sources of data can be used to create rich communication and network diagrams that can be used to understand who is talking to who in large team settings, and even more importantly, who aren't they talking to and what aren't they doing that they ought to be doing.

Event-based assessments (EBA) require substantially more preparation which partially contributes to their infrequent use (Fowlkes et al., 2009). EBAs are simulated scenarios designed to highlight the critical competencies required of a team. This technique has previously been implemented in the aviation industry to capture the various cognitive, collaborative, and advanced technological capabilities required of team personnel. These sorts of judgements add context to the complexity of assessing performance but are susceptible to reliability and validity concerns. Training raters can help alleviate these concerns (Feldman et al., 2012) but does not remove all potential biases. For example, a novice rater may be attentive and diligent in their report, but they may fail to recognize the key attributes and meta-behaviors that an expert would easily identify. On the other hand, another benefit of collecting non-invasive measures is that it does not rely upon the recall ability of the trainee and does not require as much from the instructor. Research studying cognitive workload and situational awareness frequently show us that users are often not aware of what they do not know. Subjects have difficulty recalling the workload they experienced during a task if they are asked about it much later after the exercise. Subjects are also often unable to accurately gauge their own situational awareness. The subject may believe that they were adequately surveying the situation at the time, but non-invasive measures such as eye-tracking can reveal mistakes that the trainee did not realize they made, such as failing to adequately monitor a critical gage throughout a flight task. In instructor-led after-action reviews, the debrief can often result in a he-said, she-said scenario. However, with data, instructors can point to specific moments in time where performance was not up to par and show to the trainee areas where they can improve. Non-intrusive measurements can help identify these vulnerabilities in the training environment before they cause a breakdown in performance in an operational setting. In a VR environment, performance breakdowns can be detected in-the-moment with real-time feedback given to improve the training during the scenario instead of waiting till afterwards.

**Human Operator State**

Humans unconsciously constantly give off a multitude of signals about their internal state. The subtlety of these signals ranges from those that require no training to easy to detect, such as changes in facial expressions, to more miniscule changes that are rarely detected in real-time, such as pupil dilation. Other physiological signals, such as event-related potentials (ERPs) and heart rate variability (HRV), require more complicated tools but are often quite useful when captured. These metrics are not used extensively in the military today, but all three of the human systems community of interest sub-areas highlight physiological sensors and human state measurement as a key component to achieve individualized training, human cognitive measurement for human-machine teaming, and assessment warfighter state to ensure readiness and predict health events (Defense, Science, and Technology Reliance 21, 2020).

**Eye-Tracking Data**

One of the earliest tools used by human factors psychologists to study physiological indices of trainee's performance and predict outcome was the eye tracker. Eye tracking is commonly used to assess where an operator is looking, when are they looking there, for how long, and how frequently are they checking it again. Eye tracking was previously relocated to only well-funded vision researchers, but the process and technology has become accessible enough that most college students could figure it out. This is important because more than just vision researchers are interested in eye movements. The gaze of the eye is often an indicator of the beholder's attention (Duchowski, 2007). Researchers have been capitalizing on this non-intrusive capability since the earliest beginnings of the field of human factors. Early adopters, Fitts et al., (1950), used eye tracking to study pilots' gaze during different types of standard flight maneuvers. This allowed them to collect information about what the pilots focused on during their task. Many are interested in capturing this sort of information about experts because it can be used both as a method of assessment and to help guide novices. Differences in eye-metric data have been used to assess skills level across a broad range of domains, including surgery tasks (e.g., Krupinski et al., 2006), reading tasks (e.g., Miyata et al., 2012), and driving tasks (e.g., Di Stasi et al., 2011).

Researchers interested in improving surgical outcomes studied the eye-movements of experienced surgeons with hopes to use this data to better train new operators. An analysis of similar training methodology (Tien et al., 2014) reveal the promising potential of utilizing fixation dwell time as an indicator of an operator's skill level, focus, and performance outcome on the tasks. This suggests that instructors can assess whether a trainee is ready to move on to a more difficult level of tasking based on how similar their eye-metric data matches the experts. For example, eye-tracking allows instructors to see where an operator in training is looking. Throughout the training, an instructor can assess the eye tracking data to determine whether the learner is distracted and not paying enough attention to key information within the environment. This information can help identify why a trainee is failing a task, even when the trainee may feel as if they are following all of the instructions. The instructor can then use insight from reviewing expert eye tracking data to offer guidance to the novice. This collective evidence substantially advocates for the consideration of implementing an eye-tracking component when looking to improve training outcomes.

**Heart Rate Variability**

Instructors want to be able to assess their trainees without interfering during a training scenario. In training scenarios such as air traffic control, it would be up to the instructor to directly watch the trainee to understand subtle facial, behavioral, and communication changes to understand when they start to become overwhelmed. Failing to assess a trainee while they are amid a task, results in missing data and recency bias during recall reports. Some physiological measures, such as heart rate, offer insight about an operator's state throughout a training scenario and do not impair their ability to complete the task. Researchers have shown that the heart rate fluctuates throughout the day and changes in response to physical and mental task demand. This fluctuation comes naturally and without thought; this makes heart rate an ideal contender for assessing an operator's physiological wellbeing and inferring their psychological state.

As measured, heart rate variability (HRV) is the amount of fluctuation of the length of heartbeat intervals (Malik & Camm, 1990). As a tool in research, HRV an index of an operator's workload any psychological stress. The heart's ability to vary in rate reflects the wellbeing of the operator (Acharya et al., 2006). Heart rate needs to vary to deal with the coming and passing of stressors. It is dangerous for heart rate to be consistently high or consistently low. This is not to say that a decrease in HRV is necessarily fatal or associated with dramatic decline in wellbeing. HRV is expected to temporarily decrease from time to time depending on the stressors an operator is managing. These minor fluctuations are studied in experiments and training scenarios to discern the difficulty of a task and the fitness of the operator.

Tasks of varying difficulty have been studied to better understand the capabilities of using heart rate variability as an index of workload. Early research provides evidence that rate variability decreases with increased mental workload (Mulder, 1986). Other HRV experiments have had mixed results, with some studies reporting low correlation between HRV with some self-reported measures of stress (e.g., Kageyama et al., 1998). Predominantly though, researchers have found success when using HRV as an indicator of psychological stress. HRV is a reliable indicator of workload and psychological stress in the short-term, such as during a couple of minutes, (e.g., Delaney et al., 2000; Chandola et al., 2008; Endukuru et al., 2016), as well as in assessments that measure well-being across durations longer than an hour (e.g., Kaegi et al., 1999; Uusitalo et al., 2011; Clays et al., 2011). This multitude of research suggests that HRV is a strong contender compared to other non-intrusive means of assessing an operator's state in multiple training use cases. The modern warfighter will need to be resilient with the ever-changing dynamic of the battlefield, which further highlights HRV as a beneficial unobtrusive measure to understand the warfighter's ability to handle stressful environments.

**Brain Activity**

Cardiovascular activity, pupil dilation, and other general arousal measures are useful indices of an operator's level of mental workload and stress. Some scientists rather prefer to study the source of these processes when given the opportunity, the brain itself. Electroencephalograms (EEG) and functional near-infrared spectroscopy (fNIR) are two methods of monitoring brain activity that are non-invasive nor intrusive and can be worn while an operator completes a task.

EEG devices allow scientists to assess brain voltage oscillations as they occur as discrete events. This technique is less practical for real-time training because of the data-cleaning required. This unique millisecond-temporal resolution has helped scientists assess cognitive differences in experts and novices along with the impact of balancing priorities in demanding multi-tasking environments (Strayer & Kramer, 1990; Wickens et al., 1983). fNIRS offers the ability to figuratively peer into the head of the operator, monitoring brain activity in real time. fNIRS functions by beaming an infrared light through the head's tissue that then bounces back out and is picked up a sensor adjacent to source of the light. The light that bounces back from this seemingly simple process is then analyzed to reveal information about the hemoglobin (the oxygen transporter of the blood) in that brain region (see Ferrari & Quaresima, 2012 for a historical review). This information serves as an index of mental workload and has been used in variety of training contexts. fNIRS also offers temporal resolution at the millisecond level. Their greatest appeal is their portability, as they are often no larger than a headband and are durable. The tradeoff for their portability and ease of use is that they are regionally limited compared to an EEG machine. fNIRS are typically used to collect information from a specific brain region while an EEG typically records brain activity across the entire scalp.

Studying the physiology of the human operator allows instructors to assess objective data about a trainee's performance. Newer headsets released at the time of this paper include myriads of biometric data. The *HP Omnicept* boast "data-driven insights" to support training by offering eye tracking, pupillometry, heart rate, facial recognition, and cognitive workload metrics. Other industry solutions such as the *OpenBCI Galea* headset connect with commercially available VR headsets and augment them to provide neurotechnology to VR by enabling brain activity, heart rate, and skin conductance monitoring to already sensor rich headsets that track eye-tracking. Although historic EEG systems have been infeasible to use within the training domain at scale due to the amount of calibration and noise cleaning necessary, more modern data-cleaning algorithms and the form, fit, and function of new sensor-integrated VR headsets make these new metrics a realistic possibility. These data sources provide great opportunity to leverage trainee physiological state data to understand stress, resilience, and workload. Targeting training from a multi-dimensional perspective can ensure a more comprehensive picture of readiness (Rebensky et al., 2022). This insight is often beyond what subjective measures such as self-report or behavioral observation can reveal. Biological data alone lacks the context information required to assess performance. It is for these reasons and more that instructors will often utilize a combination of performance measures. Instructors often do not need to read brain activity data to recognize that a trainee is struggling with the demands of the task. Instructors can spot these differences in an operator's interaction with the system and their interactions with the team. Augmented reality, virtual reality, and mixed reality data sources provide an opportunity to capture a wider breadth of data from the digital environment itself, input from the user which can be used to capture where they are looking and who they are interacting with, and from additional, integrated sensors such as biometric markers. Data heavy headsets such as the HP Omnicept put forth that the wealth of biometric data provided by the headset can enable adaptive extended reality (XR) training. These data sources, when used in tandem, can measure their performance as well as their cognitive state. In order to use the depth of information provided by these extensive data sources for adaptation, integrating AI and ML techniques become exceptionally valuable.

## EXAMPLES OF DATA CENTRIC VR SIMULATORS

Each of the efforts discussed below present simulators developed by the Air Force Research Laboratory's Gaming Research Integration for Learning Lab (GRILL®). The goal for the development of each simulator was to support a training or research objective to inform the Air Force on training capabilities and requirements.

### VR Allows Finite Measurement of Interactions

The first example, the parachute simulator is an immersive VR training simulator that can be configured to represent various weather conditions, times of day, and geo-typical locations. The weather conditions can be configured between each virtual "jump" to account for clear, cloudy, overcast, thunderstorms, and snowy weather. This is then combined with a selection for time of day between dawn, noon, dusk, and midnight. Finally, the environment is completed with location selected from grassy mountains, snowy mountains, or an island landscape. Each jump has a set of configurable options to make the jumps more realistic. These settings include the type of parachute used, the parachute malfunction (if any), the starting height of the jump, and when the parachute should deploy which is based on absolute altitude, delta time after jump start or delta height after jump start. Finally, each parachute type includes a set of physics parameters that are used to increase the realism of the simulator. These physics parameters can then be altered for

specific parachute malfunctions such as twisted or broken lines and canopy holes. These metrics allow for repeated exposure to different training conditions which can improve the transfer of training. Performance in each of these configurations can be tracked to understand if trainees are struggling with a particular scenario (e.g., landing performance is poor when strong winds come from behind).

**Figure 1. Parachute Simulator**

To allow for naturalistic interactions that can afford proficiency-level measurement, the simulator also utilizes physical controls in lieu of controllers. Two parachute toggles were mounted alongside the simulator to provide input from the user into the virtual environment (see Figure 1). This input is then translated into values usable by the simulation to affect the physics of simulator parachute. In addition, a customizable wind speed and direction is applied to the virtual parachute. In VR, we can track not only when the toggles are engaged, but at every second how far the trainee pulled each toggle. These measurements have the capability to capture detailed skill growth data by understanding and tracking the reaction time, smoothness, and accuracy of the toggle controls. As the final input, the simulator uses the rotation of the virtual reality headset to determine the angle of the user's eye gaze. Within VR, this information is used real-time to prompt the users to gaze at the horizon during final descent if the headset position does not detect the position to be high enough. All of these pieces of information are then used as part of the final jump summary as a set of performance outputs. The summary reports the current settings of the simulator such as location, starting height, weather, wind direction and speed, and time of day as well as parachute type and malfunction. Performance metrics are calculated based on user input and displayed to the user such as degrees off of the wind line, rate of descent, ground speed, distance from the goal location, and the angle of horizon gaze. Trainees are then able to see immediately after the scenario where they need to improve, and the instructor can spend their time coaching the trainees on how to improve their performance.

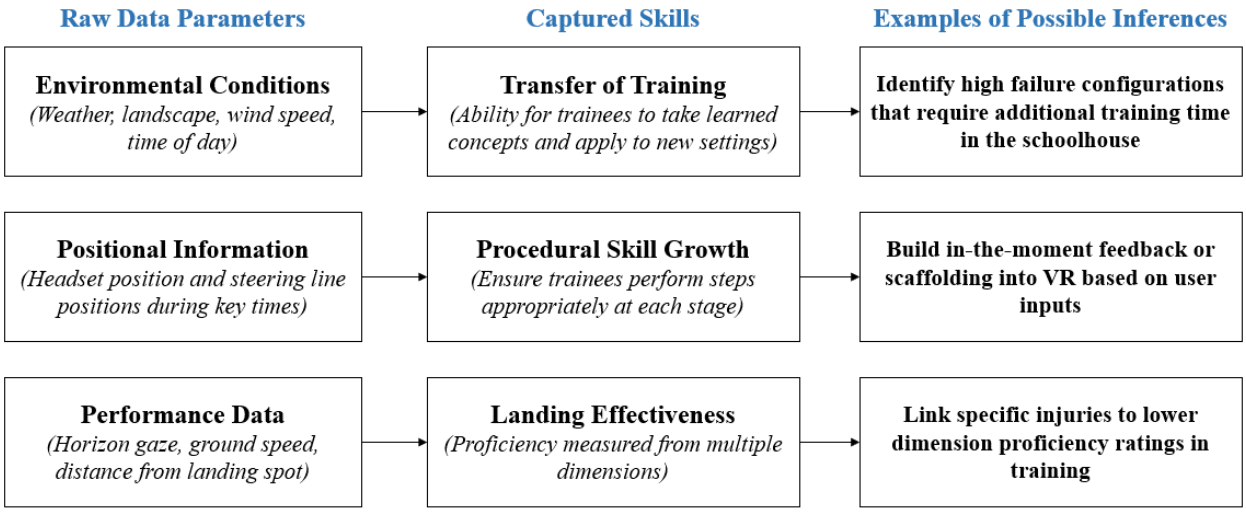


**Figure 2. Parachute Simulator Data Inputs and Outputs**

## VR Allows for Tracking Skill Growth and Trends

An important additional concept of "readiness" is supporting the health and wellbeing of our veterans and those injured in the line of duty. Data-centric VR simulators have many applications in medical training offering enhanced opportunities to gather objective data and improve trainee and patient outcomes. One such application is the development of a VR simulator designed to assist patients with Dementia and Traumatic Brain Injuries (TBI). The simulator developed through collaboration between the AFRL, Palo Alto VA Medical Center, and Naval Postgraduate School (NPS) accurately recreates the smart rooms where patients will eventually live independently. This level of realism not only enhances the patients' immersion and engagement but also facilitates the transfer of learned skills to real-life settings. The simulator, developed using Unreal Engine, creates an immersive and realistic virtual replica of

the smart rooms that these patients would eventually need to navigate and live without assistance. By incorporating data collection mechanisms within the VR environment, researchers and clinicians can gain valuable insights into the patients for personalized treatment and rehabilitation strategies. Collecting data at the task completions and interaction level lies at the core of the VR Dementia project. Within the virtual environment, patients are presented with various tasks of daily living, such as medication administration, meal preparation, and personal hygiene. As patients engage with these tasks, the simulator captures and records their interactions, response times, and performance metrics. This granular data provides valuable information about the patients' cognitive abilities, processing speed, attention, and executive function, allowing clinicians to tailor interventions and track progress over time. Providing environments such as this one, allow users a low physical demand environment can help patients re-learn and improve instrumental activities of daily living as a supplement to their rehabilitation (Greenhalgh et al., 2021). By analyzing this data, patterns and trends can be identified, enabling the identification of areas or specific tasks where patients may require additional support or specific training interventions.

The VR simulator offers a safe and controlled environment where patients can practice and refine their abilities, gaining confidence in performing ADLs and improving their quality of life. In other VR settings, creating replicas of real-world locations can allow trainees to practice before live training and refine their skills as many live training exercises have limited travel time allocated. By leveraging the power of data in VR simulations, healthcare professionals and researchers can gain valuable insights into the needs and progress of patients and trainees. The objective data collected within these simulators enables personalized treatment plans, identifies areas of improvement, and tracks patient outcomes over time. These digital twin environments can be used as a stepping stone. In medical settings, they can inform clinicians on whether a patient is ready for independent living, whereas instructors may use it to determine when the trainee's safety procedures are refined enough to practice in live settings with a low risk of injury. As data-centric VR simulators continue to evolve and integrate with AI technologies, they hold great potential in revolutionizing the field of rehabilitation and improving the lives of individuals with cognitive impairments within and outside of the military.

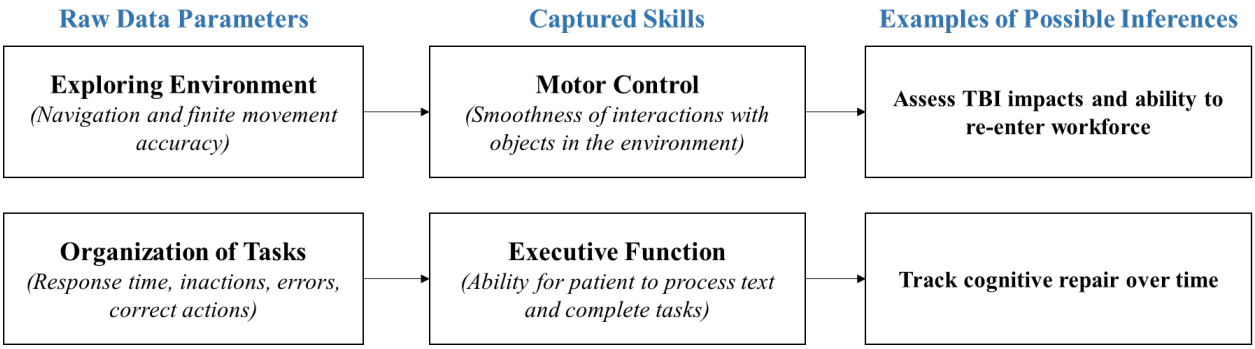


**Figure 3. VR Dementia Data Inputs and Outputs**

### VR Allows Data to Serve as a Continuous Training Loop

Utilizing not only the system and positional information but also physiological information from the trainee can provide valuable insight into the trainee's state. The Driving-Based Adaptive Research Testbed (DART) leveraged physiological sensor data and performance data in order to provide adaptive training within a three-dimensional (3D) or VR environment. DART primarily logged three kinds of data: time series physiological data, event-driven data, and summary data at the end of trials. All of these were time-stamped and synthesized within one .csv file. This allowed specific event-driven data, like an incorrect n-back response, to be viewed alongside the physiological state of the participant. Instead of a generic log where each line declares the type of data it is, the DART used a log with a unique column for every output measure which allowed for quick filtering and analysis to support after action review by instructors. Within a training environment, outputs of these data could be connected automatically to an LMS. The first data, type, performance data, can be highly granular in a VR environment. In the context of driving, performance metrics can include speed, road deviations, and any errors (e.g., having to reset or when trainees accidentally went off road). Each of these pieces of data can be highly sensitive to the proficiency of users. In the context of driving off road, it can be contextualized as (a) total time any part of the vehicle went off road, (b) total time the whole vehicle

was off road, (c) the number of times a vehicle went off road, and (d) even combinations of data such as which parts of the map the vehicle went off road or the proportion of times the vehicle went off road above a certain speed. Capturing data in this way as opposed to simple error frequencies allow us to measure proficiency of drivers over a continuum. Similar performance metrics could be captured for aviation (e.g., deviations from a flight path), and marksmanship (e.g., deviations from the target).

Within the DART environment, additional auditory tasks were layered on top of driving tasks. This allows for tracking the trainee's ability to juggle multiple tasks at any given time. As many operational settings have multiple job requirements at any given time (e.g., fly safely, communicate with the team, and identify targets), using VR data to understand the trainee's ability to meet all of these task objectives simultaneously can be beneficial for understanding a trainee's multi-dimensional proficiency. Training blocks can then be adapted based on trainee's specific KSAO deficiencies. Dynamic adaptation can optimize each trainee's learning and avoid the typical limitations of one-size-fits-all training. Ensuring that trainees are focused on developing the most relevant KSAOs delivers the greatest return on investment for training programs. This personalization, however, requires a deeper understanding of the task, data, trainee, and context. Multi-modal data sources are required in order to ensure there is a constant stream of information at a granular enough level to capture variations in behaviors that appropriately represent the cognitive state of the individual. This state is an essential element when determining the performance of the individual, which allows further training adaptation. Data must be sensitive enough to capture variability within and between trainees and the tasks must be represented with the right amount of complexity variability to appropriately capture the requirements of the task environment. Without the depth of data, live adaptations would not be possible. Additionally, with a total data tracking pipeline, on the job incidents could potentially be linked to performance during specific lessons at the schoolhouse, if it is found that trainees that underperformed on one metric have higher frequencies of incidents at their next base, it could signal back to the schoolhouse to incorporate more training time on that subject. Similarly, with more granular proficiency data, we can track over the course of training when proficiency begins to plateau. These findings can provide the opportunity to identify areas to reduce training that is already sufficiently learned and focus on deficit areas.

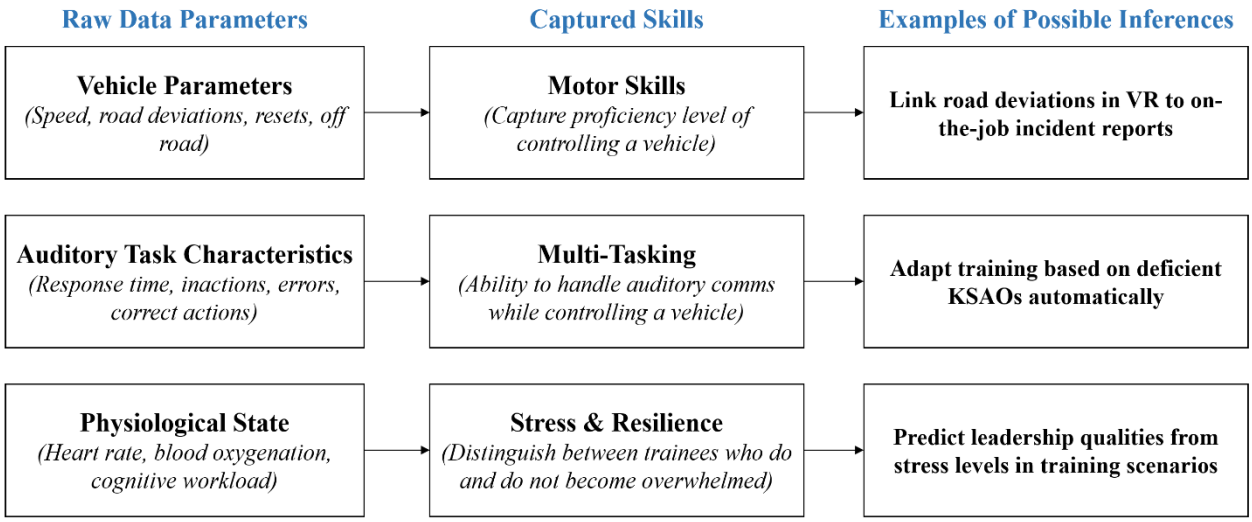


**Figure 4. DART Data Inputs and Outputs**

As for physiological metrics, the initial design of DART included connections to a Polar H10 heart rate monitor and a prototype fNIRS sensor by NIRSENSE. These metrics were included as heart rate and fNIRS can be used as predictive measures for cognitive workload (Mehler et al., 2009; Unni et al., 2005). By using these physiological metrics, an instructor could identify specific moments that resulted in high workload for trainees and review those scenarios to determine if there are particular aspects that the trainee may be struggling with. In more advanced training design, systems can be made to dynamically adapt based upon the trainee state as well as performance information. Adaptive simulators do require refinement as dynamic adaptation can be subject to error if natural fluctuations in physiological data or performance are adapted upon too frequently (Mehler et al., 2009; Rebensky et al., 2022). For DART, two-minute periods of data collection were used to track averages of change in physiological state and

performance over time, before applying an adaptation. In other training applications, two minutes is sufficiently short enough time before adapting. Simulation designers may consider utilizing newer VR head mounted displays that include sensors already built in, which can reduce any instructor needs to become familiar with sensor application. In future work upon DART, the GRILL plans to experiment with the *HP Reverb G2 Omnicept* edition which includes physiological sensors and claims cognitive load estimation.

## CONCLUSION

VR can provide extremely valuable unobtrusive and objective data collected at scale. With AI's ability to handle multiple data formats without pre-processing, data integration becomes simplified, enhancing the delivery of training content. To achieve the highest potential benefit from VR data, it must be collected at a granular level, linked to learning objectives and KSAOs, conveyed to both the trainee and instructor in timely manners, and connected to an LMS and stored in a learner profile. The community should ensure the data carries beyond the virtual environment to connect to a trainee's career. This will allow deficiencies at the edge to be identified and improved back in the virtual training. Predictive models can then be developed to indicate likelihood of trainee's success. Our presented use cases provide an overview of ways data can be used to inform readiness and lead to more robust training and readiness. Our future research aims to evaluate more sensor integrated tech, exploring brain computer interfaces, leveraging AI for interpreting data. Open areas for research and policy include understanding which types of data are best for what training contexts, as well as establishing and implementing standards for communicating training data to allow for comparison across the pipeline and forces (e.g. experience API (xAPI)). Data-centric VR training has the ability to truly allow the DoD to train like we fight and continuously improve training, but will require the community efforts, DoD, industry, academia, and research to accomplish it.

## REFERENCES

Ahuja, K., Islam, R., Parashar, V., Dey, K., Harrison, C., & Goel, M. (2018). Eyespyvr: Interactive eye sensing using off-the-shelf, smartphone-based vr headsets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 1-10.

Bailenson, J.N. (2018). Protecting nonverbal data tracked in virtual reality. *JAMA Pediatr.*

Çankaya, S. (2019). Use of VR headsets in education: a systematic review study. Journal of Educational Technology & Online Learning, 2(1), 74-88.

Chandola, T., Britton, A., Brunner, E., Hemingway, H., Malik, M., Kumari, M., ... & Marmot, M. (2008). Work stress and coronary heart disease: what are the mechanisms?. *European heart journal*, *29*(5), 640-648.

Clays, E., De Bacquer, D., Crasset, V., Kittel, F., De Smet, P., Kornitzer, M., ... & De Backer, G. (2011). The perception of work stressors is related to reduced parasympathetic activity. *International archives of occupational and environmental health*, *84*, 185-191.

Defense, Science, and Technology, Reliance 21. *Human Systems Community of Interest Roadmap 2020* [Powerpoint Slides]. Department of Defense.

Delaney, J. P. A., & Brodie, D. A. (2000). Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. *Perceptual and motor skills*, *91*(2), 515-524.

Di Stasi, L. L., Contreras, D., Cándido, A., Cañas, J. J., & Catena, A. (2011). Behavioral and eye-movement measures to track improvements in driving skills of vulnerable road users: First-time motorcycle riders. *Transportation research part F: traffic psychology and behaviour*, *14*(1), 26-35.

Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice*. Springer.

Endukuru, C. K., & Tripathi, S. (2016). Evaluation of cardiac responses to stress in healthy individuals-a non invasive evaluation by heart rate variability and stroop test. *Int J Sci Res*, *5*, 286-289.

Emerson, S., Chaparro Osman, M., Rizzardo, C., Halverson, K., Ellis, S., Anderson, A., & Haley, D. (2022). Automation and augmentation on human performance in eVTOL flight. *2022 I/ITSEC.*

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279-286.

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage*, *63*(2), 921-935.

Fitts, P. M., Jones, R. E., & Milton, J. L. (2004). Eye movements of aircraft pilots during instrument-landing approaches. *Ergonomics: Major Writings*, 56.

Floris, C., Solbiati, S., Landreani, F., Damato, G., Lenzi, B., Megale, V., & Caiani, E. G. (2020). Feasibility of heart rate and respiratory rate estimation by inertial sensors embedded in a virtual reality headset. *Sensors*, *20*(24), 7168.

Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-based approach to training (EBAT). *The international journal of aviation psychology*, *8*(3), 209-221.

Greenhalgh M, Fitzpatrick C, Rodabaugh T, Madrigal E, Timmerman M, Chung J, Ahuja D, Kennedy Q, Harris OA and Adamson MM (2021) Assessment of Task Demand and Usability of a Virtual Reality-Based Rehabilitation Protocol for Combat Related Traumatic Brain Injury From the Perspective of Veterans Affairs Healthcare Providers: A Pilot Study. *Front. Virtual Real. 2*:741578. doi: 10.3389/frvir.2021.741578

Jerome, J. & Greenberg, J. (2021). Augmented reality + virtual reality privacy & autonomy considerations in emerging, immersive digital worlds. Future of Privacy Forum.

Kaegi, D. M., Halamek, L. P., Van Hare, G. F., Howard, S. K., & Dubin, A. M. (1999). Effect of mental stress on heart rate variability: validation of simulated operating and delivery room training modules. *Pediatric Research*, *45*(7), 77-77.

Kageyama, T., Nishikido, N., Kobayashi, T., Kurokawa, Y., Kaneko, T., & Kabuto, M. (1998). Self-reported sleep quality, job stress, and daytime autonomic activities assessed in terms of short-term heart rate variability among male white-collar workers. *Industrial health*, *36*(3), 263-272.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., ... & Weinstein, R. S. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human pathology*, *37*(12), 1543-1556.

Lie, D., May, W., Richter-Lagha, R., Forest, C., Banzali, Y., & Lohenry, K. (2015). Adapting the McMaster-Ottawa scale and developing behavioral anchors for assessing performance in an interprofessional Team Observed Structured Clinical Encounter. *Medical Education Online*, *20*(1), 26691.

Malik, M., & Camm, A. J. (1990). Heart rate variability. *Clinical cardiology*, *13*(8), 570-576.

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. Transportation Research Record: *Journal of the Transportation Research Board, 2138*(1), 6–12. https://doi.org/10.3141/2138-02

Miyata, H., Minagawa-Kawai, Y., Watanabe, S., Sasaki, T., & Ueda, K. (2012). Reading speed, comprehension and eye movements while reading Japanese novels: Evidence from untrained readers and cases of speed-reading trainees. *PloS one*, *7*(5), e36091.

Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: a review. *Medical and biological engineering and computing*, *44*, 1031-1051.

Rebensky, S., Perry, S., & Bennett, W. (2022). How, when, and what to adapt: effective adaptive training through game-based development technology. *2022 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

Rosen, M. A., Weaver, S. J., Lazzara, E. H., Salas, E., Wu, T., Silvestri, S., ... & King, H. B. (2010). Tools for evaluating team performance in simulation-based training. *Journal of Emergencies, Trauma and Shock*, *3*(4), 353.

Salas, E., Wilson, K. A., Burke, C. S., & Priest, H. A. (2005). Using simulation-based training to improve patient safety: what does it take?. *The Joint Commission Journal on Quality and Patient Safety*, *31*(7), 363-371.

Schatz S, & Walcutt J. (2022). Modeling what matters: AI and the future of defense learning. *The Journal of Defense Modeling and Simulation*, *19(2),* 129-131.

Seelandt, J. C., Tschan, F., Keller, S., Beldi, G., Jenni, N., Kurmann, A., ... & Semmer, N. K. (2014). Assessing distractors and teamwork during surgery: developing an event-based method for direct observation. *BMJ quality & safety*, *23*(11), 918-929.

Spitzley, K. A. & Karduna, A. R. (2019). Feasibility of using a fully immersive virtual reality system for kinematic data collection. *Journal of Biomechanics, 87,* 172-176.

Strayer, D. L., & Kramer, A. F. (1990). Attentional requirements of automatic and controlled processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 67.

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G. Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: a systematic review. *journal of surgical research*, *191*(1), 169-178.

Unni, A., Ihme, K., Surm, H., Weber, L., Ludtke, A., Nicklas, D., Jipp, M., & Rieger, J. W. (2015). Brain activity measured with fNIRS for the prediction of cognitive workload. *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 349–354. https://doi.org/10.1109/CogInfoCom.2015.7390617

Uusitalo, A., Mets, T., Martinmäki, K., Mauno, S., Kinnunen, U., & Rusko, H. (2011). Heart rate variability related to effort at work. *Applied ergonomics*, *42*(6), 830-838.

Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: a psychophysiological analysis of the reciprocity of information-processing resources. *Science*, *221*(4615), 1080-1082.