DESIGN AND EVALUATION OF AN ADAPTIVE VIRTUAL REALITY-BASED TRAINING SYSTEM

A Thesis

by

CÉSAR IVÁN AGUILAR REYES

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee, Maryam Zahabi
Committee Members, Nancy Currie-Gregg
Camille Peres
Head of Department, Lewis Ntaimo

August 2022

Major Subject: Industrial Engineering

ABSTRACT

Successful operation of military aviation depends on effective pilot training. The current training capabilities of the United States Air Force might not be sufficient to meet the demand for new pilots. To help resolve this issue, this study focused on developing a prototype of an adaptive virtual reality (VR) training system. The system was built leveraging the three key elements of an adaptive training system including the trainee's performance measures, adaptive logic, and adaptive variables. The prototype was based on a procedure for an F-16 cockpit and included adaptive feedback, display features, and various difficulty levels to help trainees maintain an optimal level of cognitive workload while completing their training. After conducting a pilot study with 14 participants, a trend favoring the use of adaptive training was identified. Results suggest that adaptive training could improve performance and reduce workload as compared to the traditional non-adaptive VR-based training. Further work is required to further validate the findings with a larger sample size. Implementation of adaptive VR training has the potential to reduce training time and cost. The results from this study can assist in developing future adaptive VR-training systems.

DEDICATION

I dedicate this work to my mother, father, and sister, who have always supported me in any endeavor and to who I owe everything. I also dedicate this work to my wife, the newest member of my family, who always makes me become a better version of myself.

ACKNOWLEDGEMENTS

I would like to thank the Consejo Nacional de Ciencia y Tecnología (CONACYT) for their support to my graduate education. In addition, I would like to thank my fellow team members David Wozniak and Angel Ham for their contributions to this project.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Drs. Maryam Zahabi and Nancy Currie-Gregg from the Industrial and Systems Engineering Department and Dr. S. Camille Peres from the Department of Environmental & Occupational Health.

All work for this thesis was completed by the student, under the advisement of Dr. Maryam Zahabi.

Funding Sources

The funding for this study was provided by the Airforce Research Laboratory (AFRL).

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Effective pilot training is essential for the continued operations of both the transportation and defense industries. The United States Air Force's current training capabilities are not sufficient to train the 1,500 pilots per year needed to maintain their services, with a shortage of 1,925 pilots as of fiscal year 2020 (Losey, 2021). To fill this gap, technologies such as virtual reality (VR) and artificial intelligence have been used to increase the effectiveness and efficiency of training. In March 2021, the first class of pilots graduated from Undergraduate Pilot Training after seven months of training rather than the traditional twelve months (Losey, 2021). While these results are positive, further improvements are possible by new training technologies.

Virtual reality, defined as a "real or simulated environment in which a perceiver experiences telepresence" (Steuer, 1992), has been applied for education and training in various fields including healthcare (Radianti et al., 2020), defense (Bhagat et al., 2016; Pallavicini et al., 2016) and rehabilitation (Maggio et al., 2019; Rossol et al., 2011). Utilizing this technology for training is promising because of its immersive properties, ability to simulate challenging situations, and scalability.

Adaptive training has the potential to take advantage of the benefits associated with VR training. The term adaptive training is defined as "*training interventions whose content can be tailored to an individual learner's aptitudes, learning preferences, or styles prior to training and that can be adjusted, either in real time or at the end of a training session, to reflect the learner's on-task performance*"(Landsberg et al., 2010). There are three key elements for adaptive training systems including: (1) Trainee performance which is an observable characteristic that is used to gauge the trainee's success in training, (2) Adaptive logic which is a function that uses trainee performance to define the way adaptive variables change, and (3) Adaptive variables which are adjustable

features of a training task that can change the difficulty of the task based on the adaptive logic.

Adaptive training can be tailored to the user's current capabilities by using multiple performance measures (Zahabi & Abdul Razak, 2020). In this vein, adaptation can help the user maintain an optimal level of arousal, avoiding both too low and too high levels (Yerkes & Dodson, 1908). This study proposes a VR-based training system that applies the principles of adaptive training to improve pilots' training effectiveness and efficiency as compared to traditional non-adaptive training systems.

1.1 Cognitive Workload

Operator workload is a concept that refers to the load imposed on a person by a certain activity. Its definition has been tied to the demands of the activity, the effort done by the person, or the level of accomplishment of the activity (Gartner & Murphy, 1979). It is important to note that workload refers to both the physical and mental demands related to tasks, and a distinction must be made when referring to only the mental load a task may require. Cognitive workload can be defined as the degree of mental effort experienced by an individual when executing a task (Moray, 2013). Since cognitive workload cannot be directly observed, several measurement approaches have been used to estimate the mental effort required for a task. For simplicity purposes, in this work we use the term *workload* to refer to cognitive workload only, excluding any physical component of it.

There are three main categories to assess mental workload: performance measures, physiological measures and subjective ratings. In performance measures, there are two main techniques, primary task and secondary task measurement. The first uses the performance in a single task to infer workload levels, with lower performance suggesting a higher workload, however, it is possible that a person might be exerting different levels of effort to achieve the same level of performance.

This is why primary task measures work when a person is above their "red line" of workload, or the moment in which additional workload impacts performance (Grier et al., 2008). In the secondary task technique an additional task is added to the main one to aid in measurement. The secondary task is used to use-up the "spare capacity" in mental effort, and can be a simple mental arithmetic or counting exercise. Performance on the primary task is assumed to be constant, while performance on the secondary task suggests workload imposed by the primary task.

Physiological measures aim at capturing the biological responses in the body while executing a task. There is a wide variety of measurement variables and techniques to assess workload, including direct measures and derivations of cardiac activity, respiration, ocular measures, electrodermal activity, and brain signals (Charles & Nixon, 2019). Each physiological variable has a particular relationship with workload levels, and it often varies with external factors such as illumination, temperature physical activity, and others. Subjective ratings are another approach to estimate workload, including the cognitive aspect. They consist of self-reported questionnaires, given after performing a task, in which participants rate different categories related to the load imposed by the task. Examples of these techniques include the NASA-Task Load Index (Hart & Staveland, 1988) and the Modified Cooper-Harper Scale (Hill et al., 1992).

Cognitive workload is relevant in the context of training, as an appropriate level of mental effort in the adequate tasks is required to achieve positive training outcomes(Sweller, 1988). Cognitive Load Theory (CLT) (Sweller, 1988) divides cognitive load in training into three categories: intrinsic load, extrinsic load, and germane load. Intrinsic load is defined by the difficulty level or complexity of the material to be learned, corresponding to scenario difficulty level. Extrinsic load is connected to how challenging is the medium of instruction, independent of the content, which in our case is an interactive VR simulation. Germane load is the effort the trainee uses to develop

a mental model of the training material, linking the new information to past knowledge. Given these distinctions, to improve learning, extraneous load should be reduced, as it wastes available cognitive capacity, intrinsic load should be modulated, and germane load should be increased, all without exceeding the available mental capacity of the trainee.

1.2 Pilot Training

Tasks associated with operating aircraft heavily engage the pilot's psychomotor and cognitive abilities (Wise et al., 2010). One-on-one instruction is a training method that has proved to be effective in producing successful trainees as compared to traditional classroom settings (Bloom, 1984), and is used in current pilot training efforts (Hunter, 2021). However, one-on-one instruction comes with several limitations, such as high cost, restricted training capacity, and instructor-based variability in the quality of instruction(Wise et al., 2010) .

To complement this type of instruction, other technology such as the flight simulator has been used to improve the efficiency of training new pilots. This technology allows trainees to practice training objectives or scenarios repeatedly and also those that cannot be safely executed in person (Orlansky et al., 1994), such as hazardous events. Simulators also reduce cost by mitigating the need to use aircraft for training while still allowing training missions to be repeated quickly and efficiently. While these factors are helpful in reducing training costs, the effectiveness of simulators overall depends on training objectives and the phase of training (Carretta & Dunlap, 1998; Morrison & Hammon, 2000; Rogers et al., 2007; Roscoe & Bergman, 1980).

1.3 Virtual Reality

Several studies have explored possible implementations of VR into aviation related programs (Hunter, 2021; Mühlberger et al., 2001; Oberhauser & Dreyer, 2017), mostly with the purpose of

reducing training costs and increasing the speed at which new pilots can be trained (Belani, 2020). In addition to aviation, VR has been used, in the context of spaceflight, for extravehicular activity (EVA) training (Cater & Huffman, 1995; Garcia et al., 2020). Another advantage of VR is scalability, as virtual models can be replicated at lower unit cost than full-scale flight simulators(Hunter, 2021). The application of adaptive VR to further improve training for aviation has yet to gain more momentum, with few studies investigating the benefits of this type of training for pilots compared to other fields, such as health care. This type of training could improve upon the benefits of traditional VR training.

1.4 Components of an Adaptive VR-Based Training System

While research has been conducted on the benefits and costs of using VR over other training methods (Gavish et al., 2015; Schultheis & Rizzo, 2001), fewer studies focused on the effects of adaptive VR training. A literature review of the topic has found positive to mixed results on the benefits of using adaptive over traditional VR training (Zahabi & Abdul Razak, 2020). Though few studies have explored adaptive VR systems in aviation training, other fields have found that adaptive training has significantly increased performance results compared to control groups (Fricoteaux et al., 2014; Jones et al., 2016; Lang et al., 2018; Luo et al., 2013; Mariani et al., 2018; Peretz et al., 2011; Zhang & Tsai, 2021), which show a promising avenue for further research on incorporating this type of system in pilot training. The following subsections elaborate on the three main components of adaptive training and their application.

1.4.1 Performance measures

Performance measurements are used as input variables for the adaptive logic to determine how the adaptive variables should be changed. These include primary task measures, such as accuracy and

task completion time (Heloir et al., 2014; Mariani et al.; Rossol et al., 2011; Summa et al., 2015), as well as physiological measures which include heart rate, galvanic skin response, heart rate variability, and blink rate (Dahlstrom & Nahlinder, 2009; Monfort et al., 2016; Tattersall & Hockey, 1995). To apply the adaptive component, these measures have to be taken during training trials and the adaptive logic has to evaluate the results to determine the changes in adaptive variables. The methods section details how the components of adaptive training were applied to the prototype built.

1.4.2 Adaptive logic

Adaptive logic works by using methods such as rule-based procedures and classification algorithms to evaluate user performance measures, such as physiological or accuracy, and use the result to adjust the parameters of a scenario appropriately (Bian et al., 2016; Lahiri et al., 2012; Rossol et al., 2011; Saurav et al., 2018). Common classification algorithms used for this purpose make use of machine learning algorithms, such as random forests, Bayesian networks and neural networks (Zahabi & Abdul Razak, 2020). The algorithm used is generally dependent on the variables and expected outputs as well as the type of training, as there is no widely accepted standard or a single technique for adaptive training. This study explored a combined approach: a Bayesian Network as a classification algorithm, combined with a transition logic, based on the result of the classification. This method was used because of the effectiveness of the Bayesian Network as an estimator of uncertain measurements (Fenton & Neil, 2018) and its use in other adaptive training applications (Besson et al., 2013; Rossol et al., 2011).

1.4.3 Adaptive variables

When implementing adaptation into VR or simulation-based training, there are seven broad

categories of adaptive variables manipulated including: (1) the simulated environment (e.g., illumination or sound level), (2) stress or physical-based features applied to the trainee (e.g., gravity, force, and vibration), (3) controlled elements in the simulation (e.g., self-avatar), (4) the trainee's control (e.g., gain or simulated feel), (5) display features (e.g., gain, lag of the display), (6) training scenario difficulty, and (7) the secondary task load (Kelley, 1969). Of these, the most easily applicable to a pilot training task based on review of their use in previous studies are training scenario difficulty and display features (Chemuturi et al., 2013; Goettl, 1993), which were both used in this work.

Scenario difficulty refers to adjustments made to increase or decrease the effort required to complete the scenario by a trainee. Changing the difficulty level can also overlap with other adaptive variable categories. For example, changing display features, such as the speed at which stimuli are presented, can indirectly adjust the difficulty of a scenario by having trainees complete more tasks in less time. Another relevant aspect is adaptive feedback, or the modification of feedback the user receives depending on their performance, such as in the timing of feedback delivery or content(Feidakis, 2016).

1.5 Problem Statement, Research Objectives, and Hypotheses

The advent of superior performance and more accessible VR technology opens the opportunity for developing high quality training systems. In addition, adaptive training can leverage technology to deliver a better training experience in terms of effectiveness and efficiency as compared to traditional non-adaptive approaches. Therefore, adaptive VR training systems offer a high impact opportunity for occupational instruction. In the case of the USAF, the use of such a system could enhance its training capabilities and achieve pilot training requirements moving forward.

The objective of this study was to develop a prototype for an adaptive VR-based training system for pilots leveraging the three main components of adaptive training including performance measures, adaptive logic, and adaptive variables. In addition, the effectiveness of the system, in terms of its performance using a pilot test with 14 human participants, was evaluated. Based on the findings of previous studies on adaptive training (Fricoteaux et al., 2014; Lang et al., 2018; Luo et al., 2013; Mariani et al., 2018; Zhang & Tsai, 2021), it is expected that an adaptive training would improve trainees' performance more than the traditional non-adaptive VR approach.

2. METHOD

2.1 Apparatus

The VR environment was built using the Unity development platform (Unity, 2021). The computing system runs a Windows 10 System type 64-bit operating system, with an Intel(R) Core (TM) i9-10900K CPU @ 3.70GHz 3.70 GHz processor, 64 GB of RAM, and an NVIDIA GeForce RTX 3080 memory card. This system allows for high performance graphics for applications that rely on this type of functionality, such as VR systems.

The scenario was built with the HTC VIVE Pro eye VR headset as the main tool for delivery. The device was chosen for its ability to generate eye tracking data and its latency of less than 10ms in testing, a critical factor in immersive VR training (Oberhauser et al., 2018). The system was set up with two “base stations” opposite to each other, which defined the space of the VR environment. The VR equipment also consisted of a pair of controllers, which were the main way for the user to interact with the system while using the headset. Two buttons on these controllers were used to interact with the scenario as illustrated in Figure 1 below.

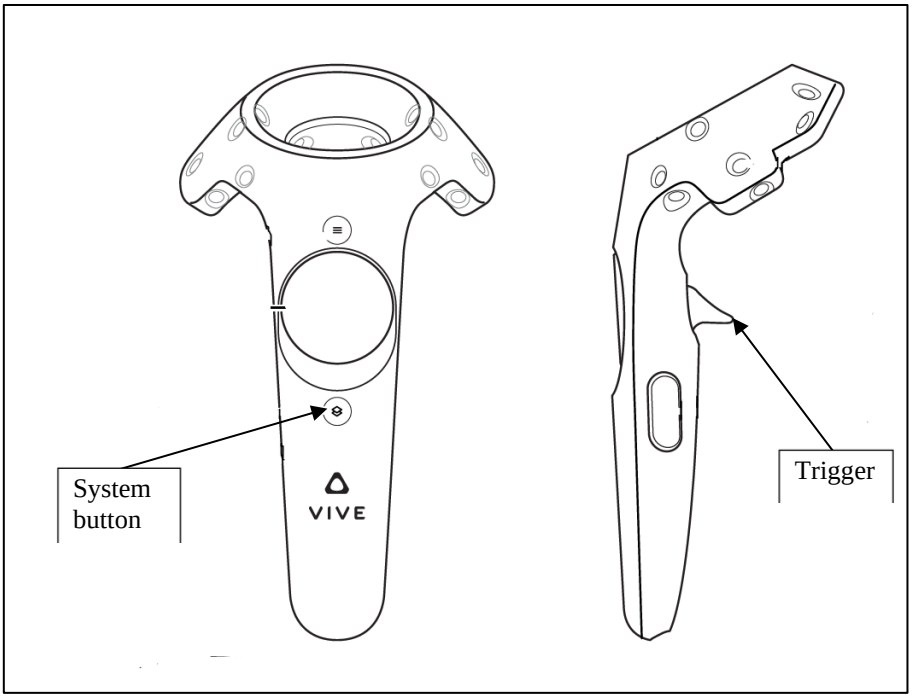

Figure 1. HTC VIVE controllers

To capture heart rate variability and electrodermal activity, we used the EMPATICA E4 heart rate monitor. The device has been verified as a reliable and valid way to measure heart rate variability and electrodermal activity (McCarthy et al., 2016; Milstein & Gordon, 2020; Schuurmans et al., 2020).

It allows for real-time data acquisition, and the data can be downloaded from a web application.

The trainee completed the scenario in a seated position next to the computer running the VR equipment. To start, the trainee wore the VIVE headset and used one of the controllers to interact with the simulation, completing the training tasks. In addition, the user fastened the EMPATICA E4 heart rate monitor on their non-dominant hand to capture the relevant data. Figures 2, 3, and 4 illustrate the equipment setup and the EMPATICA E4 heart rate monitor.

Figure 2. HTC VIVE Pro Eye headset and controller

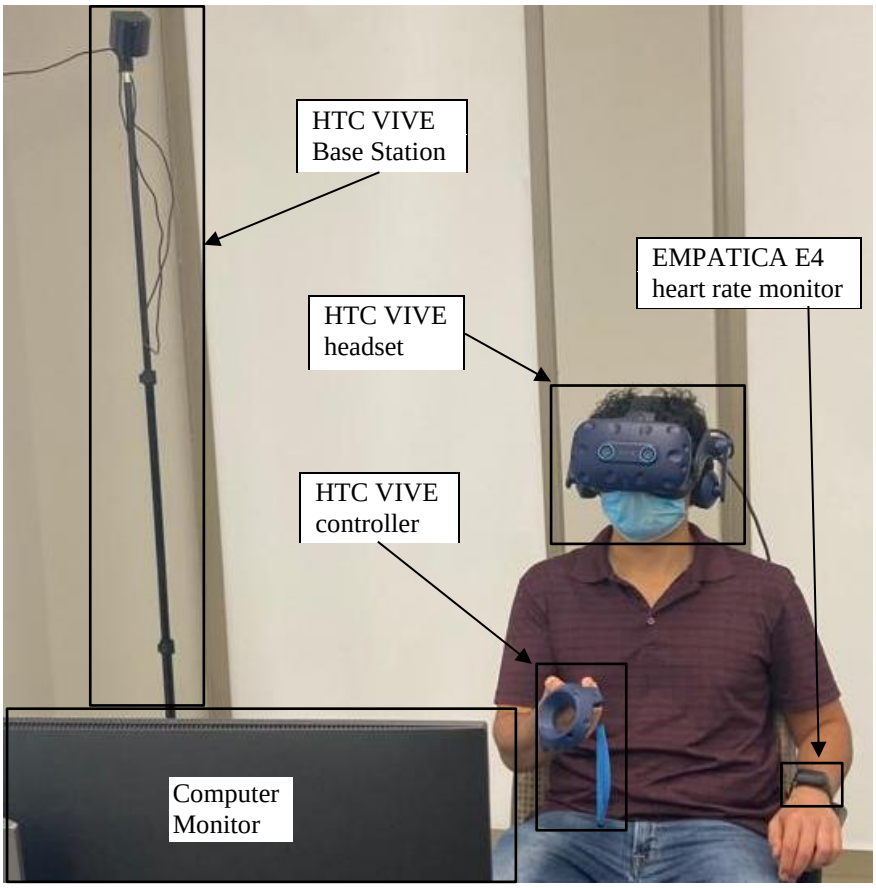

Figure 3. Study setup

Figure 4. EMPATICA E4 heart rate monitor and the data collection app

2.2 Virtual Reality Scenario

The task implemented in the scenario was based on a navigation tutorial for the F-16 fighter. The VR scenario was scripted in C# within Unity, with the adaptive component being handled by Python applications. The fighter cockpit was based on digital artwork (Moreno, 2019) and some additional elements and interactivity were added based on the scenario. The objective of the scenario was for the trainee to identify the cockpit elements in the correct order, simulating the navigation tutorial task. To achieve this, the trainee used a controller that interacted with the cockpit by hitting the trigger button. If the correct item in the procedure was hit, the checklist showed a checkmark. If incorrect cockpit elements were hit 3 times, a cross icon was displayed, and the scenario progressed to the next item in the procedure. Figure 5 displays the virtual reality cockpit from the perspective of the trainee. Note that the F-16 cockpit is visible along with three panel sections for displaying information to the user.

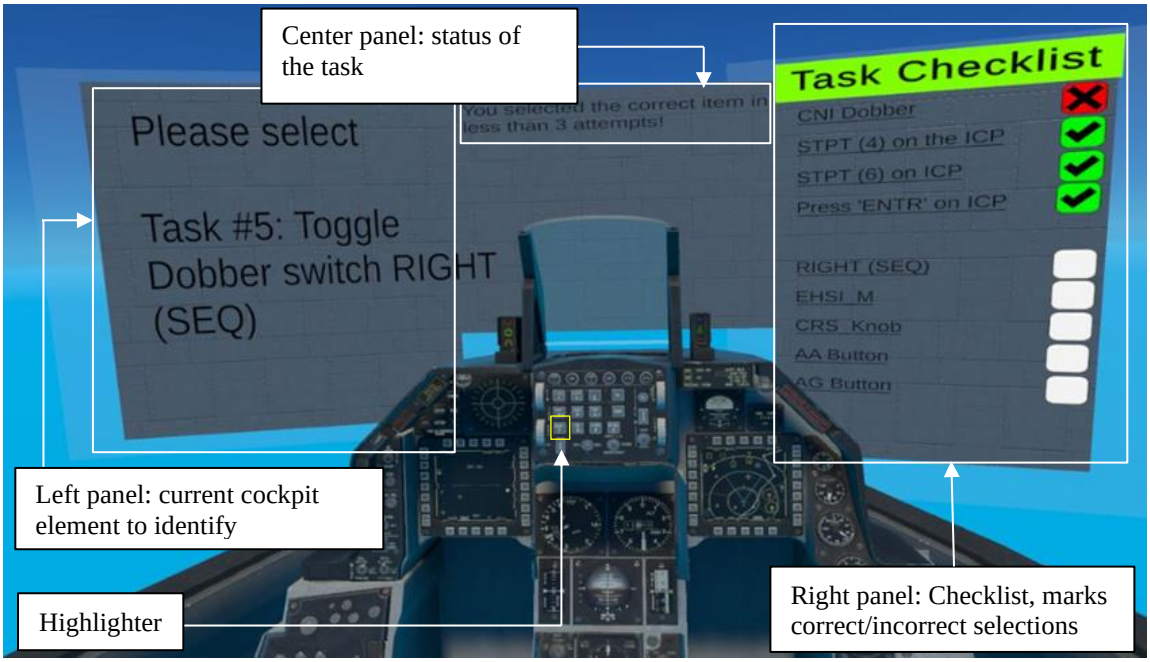

Figure 5. Virtual reality cockpit

The left panel shows the current task, and its status is displayed on the center panel. The right panel has a task checklist and informs the trainee of their performance during the scenario, with green checkmarks and red cross symbols. The trainee reads the element to be selected on the left panel, then finds the element in the cockpit. Once the element is located, the trainee uses the controller to point the laser to the element and presses the trigger to select it. After each task is completed, the program moves to the next task, looping through until the user has completed all of the tasks. At this point, the participant is assessed by the adaptive logic and will enter the next scene based on the results of the evaluation. The following section elaborates on the adaptive component of the training.

2.3 Adaptive Training Components

The proposed solution architecture for the prototype is suitable for local or networked environments. Within the local infrastructure, data is gathered using equipment, processed, and

used as input for the algorithms that make up the adaptive logic. The result of the adaptive logic is communicated to the VR infrastructure, where the training system is being implemented. The system utilizes this result to make appropriate changes in the adaptive variables, in this case, scenario difficulty and feedback. Figure 6 illustrates the solution architecture in a graphical form.

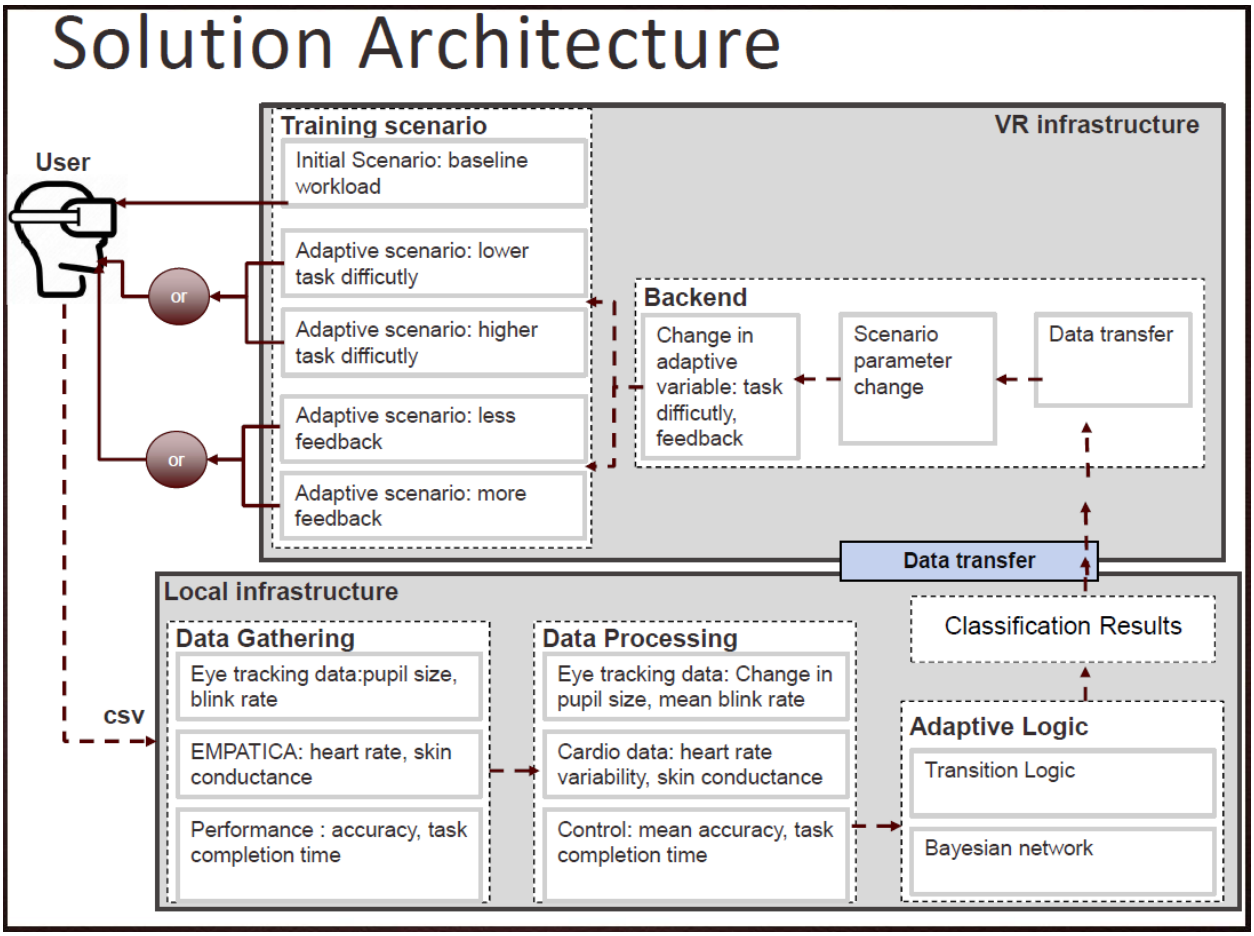

Figure 6. Adaptive VR-based training architecture

The scenario described above was designed and built as an adaptive training system prototype. As mentioned previously, an adaptive training system requires three components to operate successfully: trainee's performance measurement, adaptive logic, and adaptive feedback. The following subsections elaborate on the selection of these elements for the prototype as well as how they were implemented in the scenario.

2.3.1 Trainee performance measurement

The performance measurement in the adaptive training system was focused on trainee's estimated cognitive workload. Cognitive workload cannot be directly measured, however, there are several reliable indicators of workload that can be used to estimate it, and which were included in the system. Physiological measurements are a broad category of these indicators, the ones that were included in this prototype included heart rate variability (HRV), electrodermal activity (EDA), percentage change in pupil size (PCPS), and blink rate. Heart rate variability has been determined to be a reliable indicator of cognitive workload (Charles & Nixon, 2019) and has been used in aviation settings (Roscoe, 1992). Electrodermal activity (EDA) was also included, as it has been especially useful for sudden stimulus and is robust to repeated trials (Charles & Nixon, 2019).

Two ocular measurements for workload estimation were used in this study including PCPS and blink rate. Both are reliable indicators of cognitive workload (Charles & Nixon, 2019; Veltman, 2002). In addition to physiological measures, we recorded accuracy and task completion time as primary indicators of workload. Accuracy is defined as the total "hit points" the user scores divided by the total number of points per trial. This measure has the benefit of being an objective way to capture the effect of task difficulty on the user's mental resources assuming the participant's full attention is being devoted to the task. Task completion time was defined as the average time to complete each task from the checklist.

Trainee performance measurement was implemented in the system with a combination of methods. Heart rate variability and electrodermal activity were collected using the EMPATICA E4 heart rate monitor. The data for the two ocular measures, blink rate and PCPS, were obtained using the eye tracking capabilities of the HTC VIVE Pro Eye. Accuracy and task completion time were defined within the Unity project. For all these measurements we developed post-processing

applications in Python that transformed the original form of the data into the final form used in the adaptive logic component.

2.3.2 Adaptive logic

The adaptive logic used in this study was a Bayesian Network algorithm. This model was selected based on its adequacy in indirect estimation (Fenton & Neil, 2018) as well as its accuracy and computational speed shown in prior studies (Besson et al., 2013; Zhou et al., 2020). Bayesian models are derived from a schema of direct dependencies between a set of variables, making use of graphs and node probability tables (Fenton & Neil, 2018). There are several different types of Bayesian network schemas, and they all depend on different assumptions about the relationships between variables. For this prototype, we based the model structure on the measurement idiom, which is useful when estimating an unknown variable using different indicator variables. The model was built using the GeNIe development tool and the SMILE engine for Python integration(Druzdzel, 1999). The performance of this approach is adequate for the system, taking less than a second to run on average and more than 95% accuracy based on pilot test data.

The adaptive logic was implemented in two phases. The first was completed using Python applications, which activate after the trainee completes a scenario, using data generated by the postprocessing of the performance measurement. The data are evaluated by the Bayesian model, which yields the estimated level of workload (i.e., high, medium, low). The second phase was developed with the Unity platform and worked by reading the model result from the first phase, and used it in a transition function, which determined the next scenario to load.

There were nine scenarios with the transition logic being the following: “high” workload sets the scenario to be of a lower difficulty than the one just completed, except for the “easiest” difficulty.

This means the scenario is too difficult for the participant, and workload must be reduced. “medium” workload sets the scenario to be of the same difficulty as the one just completed, as there is enough challenge for the trainee to continue to benefit from training in this scenario. A “low” workload estimation would set the next scenario to be one of a higher difficulty, indicating the user has “mastered” the current level of difficulty. If the current scenario is at maximum difficulty, reaching a low level of workload triggers the end of the training session. Figure 7 illustrates this transition process.

A Bayesian model requires the definition of a prior probability distribution for the measurement node and node probability tables for the indicators. This allows the model to update the degree of belief, or probability, of the state of the measurement node. For workload, we defined three states of “low”, “medium”, and “high” with a uniform distribution as the prior probability, as we have no a-priori information about the trainee’s mental workload (Fenton & Neil, 2018). We defined workload level as the variable to be estimated, given an indicator variable, in this case, accuracy, task completion time, heart rate variability, percent change in pupil size, blink rate, and electrodermal activity. Their relationship is represented in Figure 8.

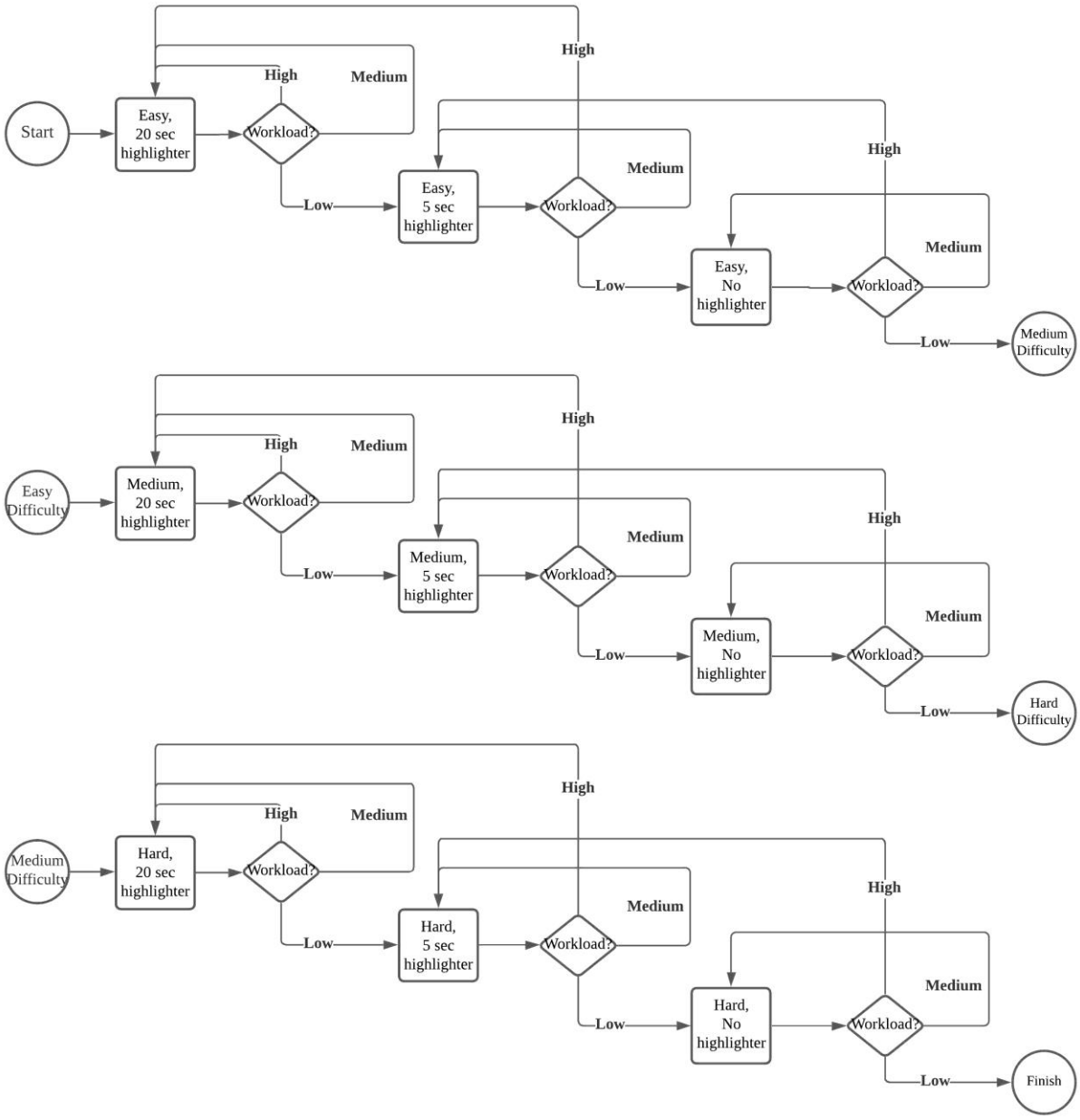

Figure 7. Transition logic

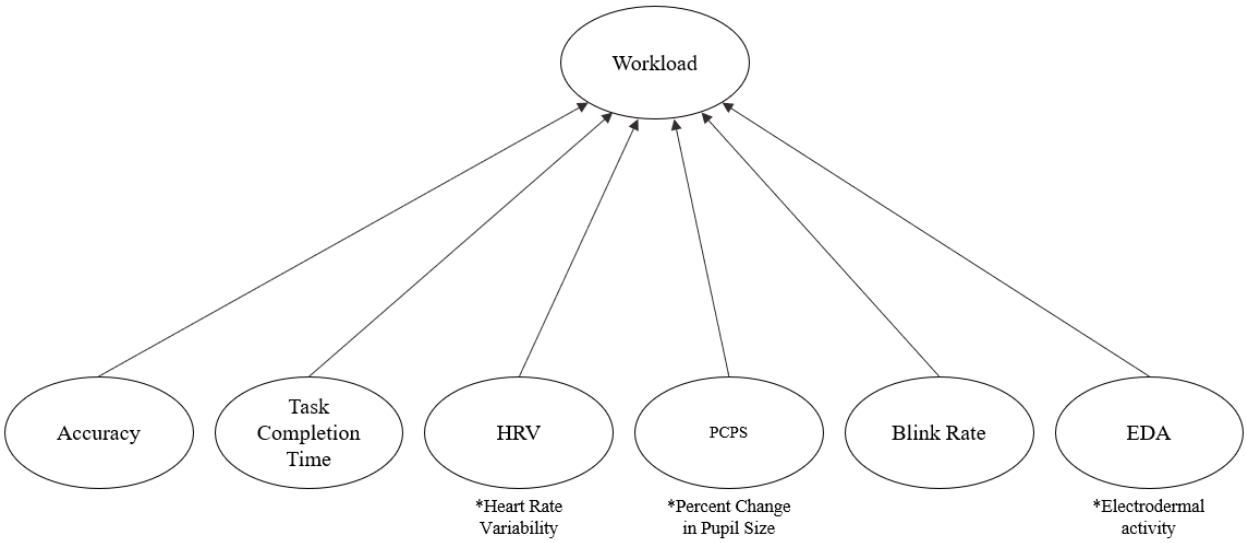

Figure 8. Bayesian network model for workload estimation

For the probability tables of each performance measurement, we defined the states based on the observed values of each variable and the relationship between these measurements and workload levels based on the literature. These models include the use of the available prior information as well as the modeler's judgement to assign a point estimate for the relationship of the indicator variable to cognitive workload. The tables also allow for a degree of uncertainty, since we assume there is no perfect indicator of workload. For instance, if a variable strongly indicates a low level of workload, we defined the probability of low workload to be 80%, however, we assigned a probability of 15% and 5% of observing medium and high workload, respectively, given the same level of the variable. In the following paragraphs, we elaborate on the node probability table definition for each performance measurement.

Accuracy and task completion time are primary task measurements of workload. As stated before, a primary task measure of workload can detect changes in workload when a person is operating near the "red line" of cognitive load, or the limit of their capacity (Grier et al., 2008), and they are fully focused on the task. This assumption was valid for this experiment due to the novelty of the

task for participants. In addition, there is a tradeoff between accuracy and task completion time. However, in this study, we did not include this tradeoff as an assumption in the model due to the fact that there was no observed relationship between accuracy and task completion time from the pilot study results. This was mainly due to the fact that, after training, the participant was faster in completing each task and their accuracy increased, contrary to what would be expected in a routine task. This can be explained by the role of learning a novel task as compared to executing a simpler or more routine task.

High accuracy levels are related to low levels of workload in aviation tasks (Gawron, 2008; NASA, 2010; Rehmann, 1995). Low levels of accuracy have been found to be related to high workload (Hicks & Wierwille, 1979; Mazloum et al., 2008). This relationship is reflected in the probability table for accuracy, shown in Table 1. As for the states, an accuracy of 50% and below was considered to be "low" accuracy, accuracy between 50% and 80% was set as "medium" and accuracy above 80% was determined as "high".

Task completion time is also a primary task measurement, which is inversely related to workload, as longer time to complete a task is associated with higher workload (Biondi et al., 2021; Mazloum et al., 2008; NASA, 2010). We defined a measurement range of task completion time that was adequate for our particular task (Rehmann, 1995). For this purpose, we tested average task completion times for different scenarios in the system and came up with the ranges of average task completion times for the node states. An average of 5.5 seconds and below was considered as "low", time between 5.5 to 6.5 seconds was defined as "medium" and any time longer than 6.5 seconds was defined as "high" task completion time. The associated node probability for workload given the task completion time state is expressed in Table 2.

Table 1

Conditional Probability Table for Workload Given Accuracy

Accuracy	Workload		
	Low	Medium	High
Low	5%	10%	80%
Medium	15%	80%	15%
High	80%	10%	5%

Table 2

Conditional Probability Table for Workload Given Task Completion Time

Task Completion time	Workload		
	Low	Medium	High
Low	60%	20%	10%
Medium	30%	60%	30%
High	10%	20%	60%

The model also utilized physiological measures of workload. A reduction in HRV is related to high workload (Jorna, 1992; Metalis, 1991; Mulder, 1992; Roscoe, 1992). This relationship has also been observed in VR settings (Hoepf et al., 2015; Labedan et al., 2021).The RMSSD method was used to generate the HRV value. As a time-domain measure, it is able to differentiate between task load levels, is sensitive to changes in task demand, and has good predictive validity for visual tasks (Charles & Nixon, 2019), making it a valid indicator for the type of task used in the study. The low level of movement experienced in the study procedure greatly reduces changes in cardiac measures do to muscular activity.

To build the node probability table for this variable, we represented the inverse relationship between HRV and workload and the fact that this measurement is more sensitive to detect changes from low levels to medium ones (Jorna, 1992; Wilson, 1992).This translates to a higher uncertainty

of workload-HRV association between medium and higher levels of workload. Therefore, the Root Mean Square of the Successive Differences (RMSSD) value of 30 and below was an indicator of "low" HRV, a value between 30-40 was "medium" HRV, and an RMSSD of 40 and above was an indicator of "high" HRV (Veltman & Gaillard, 1996; Zhang, 2007). Table 3 contains the node probability table for workload given HRV.

Table 3

Conditional Probability Table for Workload Given Heart Rate Variability

HRV	Workload		
	Low	Medium	High
Low	5%	25%	55%
Medium	15%	60%	40%
High	80%	15%	5%

Two ocular indicators of workload were used including pupil size and blink rate. Both are valid for the type of task in this study, eminently visual in nature. Although these measures are affected by ambient temperature and illumination, both variables were controlled in the study. The experiment was conducted in a room with controlled temperature, and the VR headset projected a constant illumination during the task for all participants.

Pupil size has been found to increase in higher workload conditions (Beatty, 1982; Causse et al., 2010; Kramer, 1991; May et al., 1990; Recarte & Nunes, 2000) and has been used in VR-based tasks (Abdurrahman et al., 2021; Hoepf et al., 2015).The thresholds set up for this variable state were based on the findings of prior studies (Beatty, 1982; Causse et al., 2010). To account for individual differences in pupil size, we calculated the percentage change in pupil size metric to be used in the model. A PCPS of 10% or lower was considered as "low", PCPS between 10% to 30% was set as "medium", and PCPS higher than 30% was "high" PCPS. The node probability table

for workload given PCPS is shown in Table 4. Blink rate is another ocular variable and low blink rate is an indicator of high workload (Kramer, 1991). Similar relationships have been found in the aerospace domain (Sirevaag et al., 1993; Veltman & Gaillard, 1996; Wang et al., 2016; Wilson et al., 1987) and in a VR setting (Zheng et al., 2012) . Three states of the variable were defined, with thresholds of 15 blinks per minute and less correspond to "low" blink rate, 15 to 20 blinks per minute considered as "medium", and values higher than 20 blinks per minute were grouped as "high" blink rate. These threshold values were defined based on prior studies linking blink rate levels to cognitive load (Abusharha, 2017; Brookings et al., 1996; Wang et al., 2016). The node probability table for workload given sblink rate is shown in Table 5.

Table 4

Conditional Probability Table for Workload Given Percent Change in Pupil Size

PCPS	Workload		
	Low	Medium	High
Low	80%	10%	5%
Medium	15%	80%	15%
High	5%	10%	80%

Table 5

Conditional Probability Table for Workload Given Blink Rate

Blink Rate	Workload		
	Low	Medium	High
Low	5%	10%	70%
Medium	15%	80%	25%
High	80%	10%	5%

The last physiological indicator of workload was EDA, which is known to increase as workload increases but has lower sensitivity to higher workload level. In addition, EDA is a useful method

in detecting changes in workload due to increased task demand (Charles & Nixon, 2019; Marucci et al., 2021). However, this measurement is affected by temperature, humidity, age and time of day. Of these variables, temperature and humidity were controlled in the experimental setting and the rest added to the uncertainty of the node probability table. EDA has been shown to be sensitive in capturing sudden increases in workload and is less able to discriminate between gradual changes of workload (Charles & Nixon, 2019; Collet et al., 2014). Therefore, two states of EDA were used in this study (Table 6). EDA values of 0.15μS and lower were categorized as "low" and values above 0.15μS were indicator of "high" EDA (Fairclough & Venables, 2006).

Table 6

Conditional Probability Table for Workload Given Electrodermal Activity

EDA	Workload		
	Low	Medium	High
Low	80%	40%	30%
High	20%	60%	70%

2.3.3 Adaptive variables

Another component of an adaptive training system is the adaptive variables. These variables depend on the result of the adaptive logic. In this study, the two adaptive variables were the scenario's difficulty level and display features. Difficulty level was defined by the task load of each scenario, determined by the number of tasks that needed to be completed, with more tasks meaning a higher difficulty. Higher task load has been associated with higher workload (Colle & Reid, 1998), and number of tasks has been used as a way to determine task load in several studies (Backs et al., 2000; Colle & Reid, 1998; Wilson & Russell, 2003).

Display features included a highlighter to help the user in locating the different relevant elements

in the scenario. If the user exhibited low level of workload, a scenario with more difficulty was assigned. Conversely, an overwhelming cognitive load would mean lower levels of difficulty are presented. If the trainee exhibited low levels of workload, no highlighter was shown. Moderate to high workload levels corresponded to a short (5 second) and longer (20 second) highlighter durations respectively. Difficulty level and highlighter generated nine possible difficulty pairs, presented in Table 7.

Table 7

Difficulty level combinations for the adaptive training system

Task Difficulty	Highlighter	Scenario difficulty pair
Easy	Highlighter - 20 sec	Easy difficulty, 20s highlighter
	Highlighter - 5 sec	Easy difficulty, 5s highlighter
	Highlighter - none	Easy difficulty, no highlighter
Medium	Highlighter - 20 sec	Medium difficulty, 20s highlighter
	Highlighter - 5 sec	Medium difficulty, 5s highlighter
	Highlighter - none	Medium difficulty, no highlighter
Hard	Highlighter - 20 sec	Hard difficulty, 20s highlighter
	Highlighter - 5 sec	Hard difficulty, 5s highlighter
	Highlighter - none	Hard difficulty, no highlighter

2.4 Study Design and Variables

To evaluate the effectiveness of the adaptive training system, a between-subject, pre-test/post-test experiment was designed. The independent variable was the training type with two levels including adaptive and non-adaptive VR training systems. In the non-adaptive VR training system, a monotonic fixed progression between difficulty levels was defined, independent of the results of the adaptive logic. Both system options included pre-test and post-test scenarios of identical

difficulty level, so the performance responses can be compared. The pre-test scenario was completed before the training started, and the post-test was presented after the training phase was completed. In addition, the participant completed a retention scenario, with the same difficulty as pre-test/post-test scenarios, one week after the first training session was completed. The dependent variables included accuracy, task completion time, and workload during the pre-test, post-test, and retention scenarios. Accuracy was defined as the total "hit points" the user scores divided by the total number of points per trial. Task completion time was defined as the average time to complete each task from the checklist in the scenario. Workload was defined as the workload estimated through the Bayesian model deployed in the adaptive logic.

2.5 Participants

A pilot test was conducted with 14 healthy participants (7 males, 7 females) within the range of 22 to 31 years (M= 25.5 yrs., SD= 2.7 yrs.). The participants were recruited from student population at Texas A&M university who were interested in aviation or virtual reality. All participants had normal or corrected-to-normal vision with contact lenses, with five participants using contacts. Also, they did not have any history of simulator sickness during VR use, and none exhibited simulator-induced motion sickness during or after the procedure. Two participants had previous experience piloting an aircraft. All participants read and signed the informed consent form prior to participating in the experiment. The Texas A&M University Institutional Review Board (IRB) approved the study protocol. Participants were compensated $30 for their time.

2.6 Procedure

Upon arrival to the lab, participants were randomly assigned to either the adaptive or non-adaptive VR-based training. Once the participant read and signed the consent form, they fastened an

EMPATICA E4 heart rate monitor to their wrist to measure electro-cardiac activity, and then proceeded to fill out the demographic questionnaire. Subsequently, the participant was presented with training material regarding the scenario to be performed, which consisted of short slide presentation on a computer. The participant was evaluated on the knowledge of the cockpit elements attained in the training via a paper-based examination, which involved relating a cockpit element name with its correct position in the cockpit. Emphasis was placed on correct identification of cockpit elements, as opposed to fast performance.

The simulator sickness questionnaire (SSQ) was used to identify motion sickness symptoms that could negatively affect participants prior and during the experiment (Kennedy et al., 1993). The participant put on the HTC VIVE Pro Eye virtual reality headset to engage in the virtual reality simulation. The experiment started with a "baseline" data collection phase, in which the participant relaxed within the virtual reality environment and eye tracking and heart rate data were collected for duration of 2 minutes. After that, the participant proceeded to the next phase, which was the pre-test scenario corresponding to the highest difficulty level with no highlighter. After that, the participant executed the training scenarios, which include the different difficulty levels described in the earlier section. In the non-adaptive training scenario, the difficulty level increases by one level after each trial, covering the nine possible combinations. In adaptive training scenarios, the participant started with an appropriate difficulty level (defined by the adaptive logic result for the pre-test scenario), and the difficulty level increased or decreased according to the adaptive logic, for a maximum of 18 trials, or until the participant completed the hardest difficulty. Upon completion of the training phase, the participant moved to the post-test scenario, which was identical to the pre-test scenario in terms of the difficulty level to compare performance. Upon completion of the post-test scenario, the participant was debriefed and left the lab. After one week

of the initial experiment, the participant returned to complete a retention phase, consisted of a retention scenario. This procedure was similar for both adaptive and non-adaptive experimental groups, with the only difference being the transition logic in the training phases, which is adaptive for the first and fixed for the second, in the manner explained in the previous subsection. The study procedure is depicted in Figure 9.

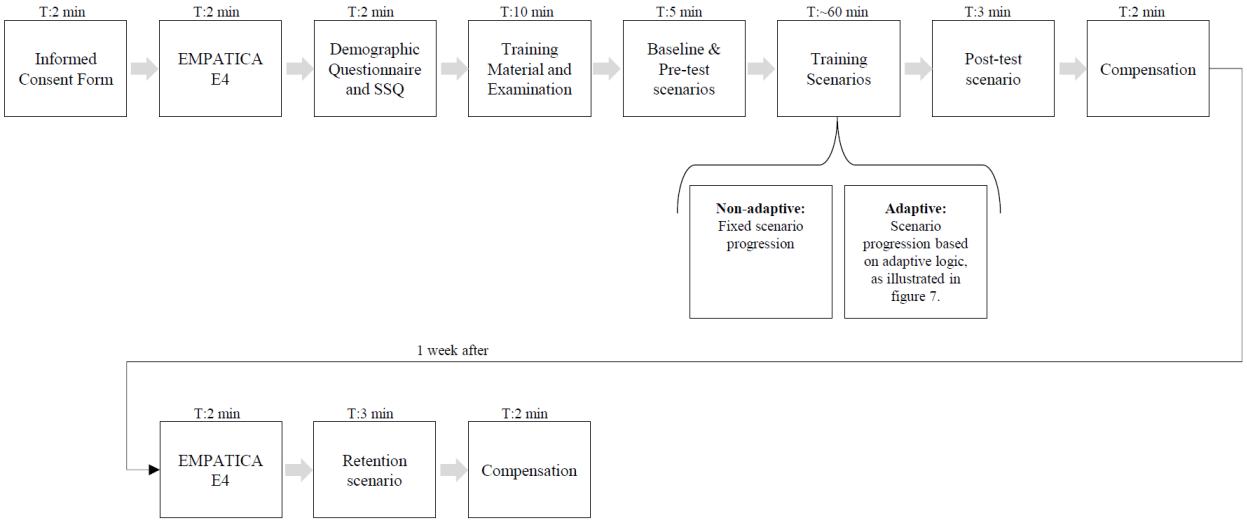

Figure 9. Study Procedure

2.7 Data Analysis

The data for the system was extracted with three main components including the eye tracking capability of the HTC VIVE Pro Eye headset, the EMPATICA E4, and the VR scenario code in Unity. The data collected from the HTC VIVE Pro Eye headset was extracted using the *VIVE SRanipal SDK* and customized code in C#. With this data, blink rate and PCPS were calculated via Python modules. The EMPATICA E4 data was obtained with the help of the *E4 Realtime* mobile application and downloaded from the *E4 connect* web service. The raw data was processed via Python modules to compute EDA and HRV (using the RMSSD method). Accuracy and task completion time were generated automatically by the Unity project after each scenario ended.

There dependent variables were analyzed: accuracy, task completion time, and estimated workload, which were measured for the pre-test, post-test, and retention phases. Accuracy was defined as the total "hit points" the user scores divided by the total number of points in each scenario. Task completion time corresponded to the total duration of the scenario divided by the number checklist items, thereby being the average task duration per item. Estimated workload corresponds to the result of the Bayesian model, used in the logic in adaptive training, and calculated, but not used for transition, in the non-adaptive version of training. To evaluate the effect of training type, the mean percent difference between the pre-test and post-test phases, and between pre-test and retention were calculated. This metric allowed us to compare the change in accuracy and task completion time for each treatment group independent from the initial level of performance. In the case of task completion time, the absolute value of the reduction in task completion time was illustrated. The higher the absolute value of the reduction in task completion time, the more effective the training method was.

To analyze the change in estimated workload, we used the reduction in workload for each participant for the three evaluation scenarios: pre-training, post-training and retention. If the estimated workload was lower in post-training or retention than in pre-training, the variable "reduction in estimated workload" was classified as "true", otherwise as "false". For instance, if a participant went from a "High" workload evaluation in pre-training to either a "Medium" or "Low" estimation, it was classified as a workload reduction instance. If the participant went to a higher or same estimate, they would be classified as no workload reduction.

The statistical analyses were performed in R(R Core Team, 2018). Before statistical testing, a screening process was conducted to identify outliers in the data, none of which were found. Subsequently, diagnostics were conducted to check for assumptions normality (using Q-Q plots

and Shapiro-Wilk normality test) and variance homoscedasticity (using the Levene test) prior to conducting One-Way Repeated Measures Analysis of Variance (ANOVA). The significance criterion of the study was set at $p \leq 0.05$ level.

3. RESULTS

3.1. Performance

The preliminary results showed that adaptive training led to a higher mean percent change in accuracy than the non-adaptive training in both pre-test vs. post-test (53.5% vs 25.4%) and pre-test vs. retention (49.0% vs 18.1%) comparisons. It is important to point out that due to the limited sample size, this observation should be regarded as a trend (and not a significant difference) that favors adaptive training. Adaptive training also exhibited a higher reduction in time in pre-test vs. retention (67.1% vs 59.8%) and a lower one in pre-test vs. post-test (74.5% vs 82.1%). Figures 10 and 11 illustrate the mean percent change in accuracy and the mean percent reduction in task completion time respectively.

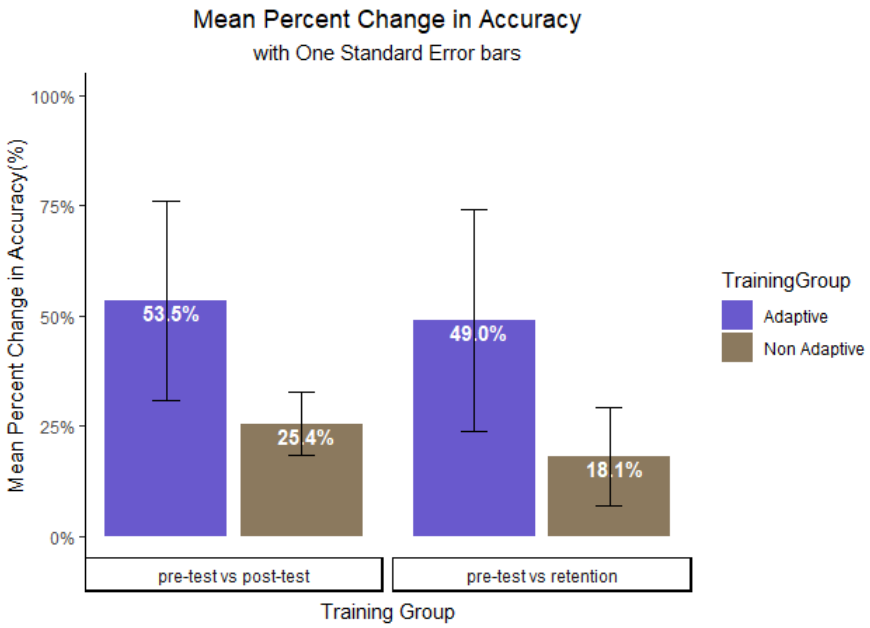

Figure 10. Mean Percent Change in Accuracy

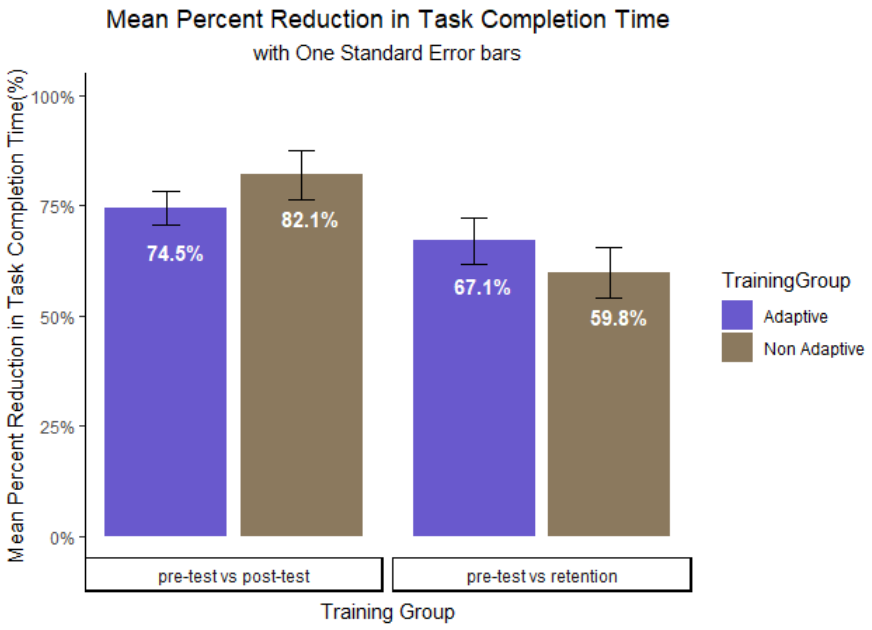

Figure 11. Mean Percent Reduction in Task Completion Time

We conducted a One-Way Repeated-measures ANOVA test to assess the significance of the differences in change in accuracy and reduction in task completion time for the adaptive and non-adaptive groups. After performing the tests using the R package (R Core Team, 2018), we did not find any significant differences in mean percent change in accuracy and mean percent reduction in task completion time. The results are summarized in Table 8.

Table 8

Results of the repeated-measures ANOVA for comparing the adaptive and non-adaptive training Groups

Dependent Variable	Adaptive training		Non-adaptive training		F	*p*	*η2*
	M	SE	M	SE			
	Pre-test vs Post-test						
Mean Percent Change in Accuracy	53.5%	22.6%	25.4%	7.2%	1.2	0.295	0.09
Mean Percent Reduction in Task Completion Time	74.5%	3.9%	82.1%	5.5%	1.08	0.318	0.08
	Pre-test vs Retention						
Mean Percent Change in Accuracy	48.9%	25.3%	18.8%	11.3%	1.21	0.293	0.09
Mean Percent Reduction in Task Completion Time	67.1%	5.2%	59.8%	5.8%	0.75	0.402	0.06

3.2. Workload

In addition to performance comparison, the reduction in estimated workload for both training groups was measured. Table 9 illustrates the reduction in workload for both training groups. We can observe a higher reduction in estimated workload in the adaptive training group in both pre-test vs post-test (85.7% vs 28.6%) and pre-test vs retention (71.4% vs 28.6%). Similar to accuracy and task completion time results, in adaptive training, we observe a higher reduction in estimated workload immediately after performing the training session than one week later.

Table 9

Reduction in Estimated Workload

Phase Comparison	Reduction in Estimated Workload	
	Adaptive	Non-Adaptive
Pre-test vs Post-test	85.7%	28.6%
Pre-test vs Retention	71.4%	28.6%

4. DISCUSSION

A prototype of an adaptive VR-based training system aimed at training pilots was developed using the Unity platform and Python as software tools and the HTC VIVE Pro Eye and EMPATICA E4 as hardware tools. This prototype was built leveraging the three fundamental components for an adaptive training system including trainee's performance measurement, adaptive logic, and adaptive variables. The system described in this work is a proof of concept, which confirms the possibility to build an interactive, real-time adaptive training system in a VR environment for pilots using a combination of hardware and open software. We consider it as a valuable starting point to either refine or build new adaptive training systems using VR. Researchers can utilize the different approaches detailed in this work to inform their development process. This prototype system was able to integrate different components and technologies to achieve the desired functionality, enabling flexibility in future efforts by not depending on an off-the-shelf platform or paradigm.

In addition, a preliminary study was conducted to evaluate the developed system. The findings suggested that adaptive training was better in terms of increasing the accuracy of trainees in both pre-test vs. post-test and pre-test vs. retention comparisons. This result should be considered a trend (and not a statistical significance) due to the limited sample size. Prior studies also observed a similar pattern of results (Fricoteaux et al., 2014; Landsberg et al., 2012; Luo et al., 2013; Mariani et al., 2018; Zhang & Tsai, 2021) which indicates that an adaptive training approach can be more effective than a traditional approach in improving trainees' accuracy. Since these findings were captured from larger sample sizes (ranging from 10 to 40 participants per group), it is reasonable to expect that the trend found for the system developed in this study correspond to the initial hypothesis but the findings should be confirmed with a larger sample size.

In terms of task completion time, there was a similar reduction for the adaptive and non-adaptive groups in the pre-training vs. post training, but a greater reduction in pre-training vs. retention for the adaptive training approach. These findings can be due to the fact that immediately after the training session, participants in both groups are familiar with the scenario and are able to go through it at a similar speed. In contrast, a greater reduction in task completion time a week after may indicate better performance in knowledge retention for adaptive training. These results may indicate a trend favorable for retention of the learned skills in the adaptive training approach. The findings are similar to the results of previous studies (Lang et al., 2018; Rossol et al., 2011) that link performance improvements in terms of time with adaptive training. However, a study with a larger sample size is required to validate these findings.

Results also suggested that adaptive training was more effective in reducing workload than the non-adaptive training group in both pre-test vs. post-test and pre-test vs. retention comparisons. Evidence regarding the effect of adaptive VR training on workload is sparse. Ariali and Zinn (2021) compared the level of perceived workload between adaptive and non-adaptive VR training systems and found no difference between the two approaches (Ariali & Zinn, 2021). It is important to note that the estimate of workload in this work is closely related to accuracy and task completion time findings. Since accuracy was improved for the adaptive training group in general and task completion time was reduced in the retention session, we can infer that the reduction in workload was influenced by the performance improvement observed for the adaptive training group. We consider understanding the effect of adaptive training on cognitive workload a relevant avenue for future research.

The results obtained for adaptive training can be attributed to several distinctions present between the adaptive and non-adaptive training approaches. For those participants that showed high or

medium workload, the adaptive logic enabled more practice to get familiar with the cockpit elements. However, adaptive training placed individuals with low estimated workload in the pre-test scenario to start at a commensurate level of difficulty, as opposed to the non-adaptive approach where all trainees started at a same level. This resulted in experiencing fewer number of scenarios for participants showing low initial workload estimates in adaptive training, with 6 scenarios on average, as compared to trainees with high initial workload, with an average of 11 scenarios, while the non-adaptive group experienced a fixed number of 9 scenarios. The fact that participants with higher workload levels at the onset of training completed, on average, a higher number of scenarios than the non-adaptive group, means adaptive training did assign participants to the right number of iterations given their workload levels.

In the context of CLT, we can say that adaptive VR based training, extrinsic load is low and constant, because VR tries to simulate a realistic environment and reduces external distractions due to its immersive properties. Intrinsic load is modulated via the adaptive component, avoiding mental overload by basing scenario transition on the estimated workload of the participant. Germane load is promoted by enabling trainees to use their cognitive resources to create an accurate mental model of the task. All of these characteristics conform to a training approach that fosters better learning outcomes according to CLT, further supporting the use of adaptive VR systems for training.

The approach used in this study has the potential to improve pilot training in terms of efficiency and effectiveness. For efficiency, we consider that the cost to set up a training system similar to the one developed is lower, on a unit basis, than traditional flight simulators, and, conversely, to the use of aircraft. In addition, adaptive training was found to potentially be effective in reducing trainees' workload and improving task accuracy. VR systems can also be effective in initial phases

of training and free up the use of other training resources such as simulators, instructors, and aircraft for later stages or more complex trainings. This can also mean greater effectiveness, as trainees can make better use of the aforementioned resources.

5. LIMITATIONS AND FUTURE WORK

This study had some limitations that must be taken into consideration. One is the prototypical nature of the system, which influenced some design choices. The approach used for building the system is sound because it was done in an efficient fashion, including all the fundamental elements of an adaptive training system. The limitations come in terms of the immersive and aesthetic qualities of the environment built, such as including more interactive controls, embellishment of the cockpit, and other features such as sound, all requiring time and specialized personnel to develop. While a more immersive system can improve the user experience, we consider that the system built achieved the main task, which was to prove the feasibility of building a VR-based adaptive training system and evaluating its effectiveness in a pilot study. To further advance towards the latter goal, a larger sample size is required to verify the trends identified in this study. To achieve a test with a power of 80% at the 0.05 significance level, and to distinguish a 20% difference in means, we would need a sample size of at least 10 participants per group. Currently, we are working on recruiting more participants for the study to reach the required sample size and be able to conduct inferential statistical tests.

An important limitation for this system is the type of the simulated task (i.e., a checklist task instead of a flight simulation task). This could reduce the generalizability of results. Due to the prototypical nature of the system, some of the necessary components to simulate flight tasks, such as flight physics, full cockpit “functionality” simulation, and relevant environmental elements such as elevation and runways were beyond the scope of the project. Regarding the adaptive logic, we

used node probability tables based on different workload indicators. While this approach was informed by relevant literature concerning the relationship between cognitive workload and each indicator variable, no standardized method exists to precisely quantify this relationship. As a result, the output of the model must be considered an estimation of cognitive workload. Future studies should explore the use of fuzzy logic (Zadeh, 1978) or ranked nodes (Fenton et al., 2007) to incorporate uncertainty into each indicator node. Both methods allow for eliciting full probability tables from limited information, avoiding the use of point estimates.

An interesting avenue of future work is adjusting other aspects of the adaptive components of the system, such as the adaptive logic. For instance, we can experiment the effects of a more dynamic adaptive logic vs. a more conservative one. This could establish an adaptive logic that reduces the total time spent on training while maintaining good outcomes in terms of performance, leading to increased training efficiency. In addition, different classes of adaptive variables could be proposed, such as various environmental features (i.e. illumination, noise). There is a need for manipulating different types of adaptive variables and identifying those that are associated with greater increases in performance.

More studies need to be conducted to validate the effectiveness of adaptive VR training in aviation. Our design decisions used the best information available in the current literature, however, the evidence of the effectiveness of adaptive training for pilots in a VR environment is sparse. As found in Losey (2021), there is a wide practical application of VR in pilot training and using appropriate adaptive components can improve training effectiveness. A benefit of the system presented is that it provides a novel platform to initiate research on evaluating the effectiveness of the adaptive VR-based training systems in aviation and other fields.

Another avenue for future research is to increase the fidelity of the VR environment to improve

the user experience of the trainee. Better graphical embellishments, interactive "mechanical" effects, audio, and adding a more realistic environment are all elements that improve the fidelity of the system. Finally, this study can serve as a platform to initiate the development of more advanced adaptive VR-based training systems.

6. CONCLUSION

Adaptive VR training has the potential to increase the effectiveness and efficiency of pilot training, giving the USAF an opportunity to enhance its training capabilities. The objectives of this study were to develop an adaptive VR training system for pilots and evaluate its effectiveness in improving training outcomes. An adaptive VR training system was developed using open software and based on the fundamental principles of adaptive training. Scenario difficulty and feedback were modified as adaptive variables using an estimate of workload derived from an adaptive logic based on Bayesian networks. This logic estimated workload using physiological and task performance data. The findings of our pilot study illustrated a trend that adaptive training could be better at improving performance and reducing workload than the traditional non-adaptive VR training. Future studies can utilize the work presented as a guideline for developing other VR adaptive training systems, which could incorporate more immersive features to improve the instructional experience.

REFERENCES

Abdurrahman, U. A., Yeh, S.-C., Wong, Y., & Wei, L. (2021). Effects of Neuro-Cognitive Load on Learning Transfer Using a Virtual Reality-Based Driving System. *Big Data and Cognitive Computing, 5*(4), 54.

Abusharha, A. A. (2017). Changes in blink rate and ocular symptoms during different reading tasks. *Clinical optometry, 9*, 133.

Ariali, S., & Zinn, B. (2021). Adaptive training of the mental rotation ability in an immersive virtual environment.

Backs, R. W., Navidzadeh, H. T., & Xu, X. (2000, 2000). Cardiorespiratory indices of mental workload during simulated air traffic control.

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

Belani, M. (2020). Evaluating virtual reality as a medium for vocational skill training. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems

Besson, P., Bourdin, C., Bringoux, L., Dousset, E., Maïano, C., Marqueste, T., Mestre, D. R., Gaetan, S., Baudry, J.-P., & Vercher, J.-L. (2013). Effectiveness of physiological and psychological features to estimate helicopter pilots' workload: A Bayesian network approach. *IEEE Transactions on Intelligent Transportation Systems, 14*(4), 1872-1881.

Bhagat, K. K., Liou, W.-K., & Chang, C.-Y. (2016). A cost-effective interactive 3D virtual reality system applied to military live firing training. *Virtual Reality, 20*(2), 127-140.

Bian, D., Wade, J., Warren, Z., & Sarkar, N. (2016, 2016). Online engagement detection and task adaptation in a virtual reality based driving simulator for autism intervention.

Biondi, F. N., Cacanindin, A., Douglas, C., & Cort, J. (2021). Overloaded and at work: Investigating the effect of cognitive workload on assembly task performance. *Human Factors, 63*(5), 813-820.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher, 13*(6), 4-16.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology, 42*(3), 361-377.

Carretta, T. R., & Dunlap, R. D. (1998). *Transfer of Training Effectiveness in Flight Simulation: 1986 to 1997*.

Cater, J. P., & Huffman, S. D. (1995). Use of the Remote Access Virtual Environment Network (RAVEN) for Coordinated IVA—EVA Astronaut Training and Evaluation. *Presence: Teleoperators & Virtual Environments, 4*(2), 103-109.

Causse, M., Sénard, J.-M., Démonet, J. F., & Pastor, J. (2010). Monitoring cognitive and emotional processes through pupil and cardiac response during dynamic versus logical task. *Applied psychophysiology and biofeedback, 35*(2), 115-123.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics, 74*, 221-232.

Chemuturi, R., Amirabdollahian, F., & Dautenhahn, K. (2013). Adaptive training algorithm for robot-assisted upper-arm rehabilitation, applicable to individualised and therapeutic human-robot interaction. *Journal of neuroengineering and rehabilitation, 10*(1), 1-18.

Colle, H. A., & Reid, G. B. (1998). Context effects in subjective mental workload ratings. *Human Factors, 40*(4), 591-600.

Collet, C., Salvia, E., & Petit-Boulanger, C. (2014). Measuring workload with electrodermal activity during common braking actions. *Ergonomics, 57*(6), 886-896.

Dahlstrom, N., & Nahlinder, S. (2009). Mental workload in aircraft and simulator during basic civil aviation training. *The International journal of aviation psychology, 19*(4), 309-325.

Druzdzel, M. J. (1999, 1999). SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: a development environment for graphical decision-theoretic models.

Fairclough, S. H., & Venables, L. (2006). Prediction of subjective states from psychophysiology: A multivariate approach. *Biological psychology, 71*(1), 100-110.

Feidakis, M. (2016). A review of emotion-aware systems for e-learning in virtual environments. *Formative assessment, learning data analytics and gamification*, 217-242.

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Fenton, N. E., Neil, M., & Caballero, J. G. (2007). Using ranked nodes to model qualitative judgments in Bayesian networks. *IEEE Transactions on Knowledge and Data Engineering, 19*(10), 1420-1432.

Fricoteaux, L., Thouvenin, I., & Mestre, D. (2014). GULLIVER: a decision-making system based on user observation for an adaptive training in informed virtual environments. *Engineering Applications of Artificial Intelligence, 33*, 47-57.

Garcia, A. D., Schlueter, J., & Paddock, E. (2020, 2020). Training astronauts using hardware-in-the-loop simulations and virtual reality.

Gartner, W. B., & Murphy, M. R. (1979). CONCEPTS OF PATTMEO. *Survey of Methods to Assess Workload*, 3.

Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments, 23*(6), 778-798.

Gawron, V. J. (2008). *Human performance, workload, and situational awareness measures handbook*. Crc Press.

Goettl, B. P. (1993). Analysis of skill on a flight simulator: Implications for training. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., & St. John, M. (2008, 2008). The red-line of workload: Theory, research, and design.

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

Heloir, A., Nunnari, F., Haudegond, S., Havrez, C., Lebrun, Y., & Kolski, C. (2014, 2014). Design and evaluation of a self adaptive architecture for upper-limb rehabilitation. ICTs for Improving Patients Rehabilitation Research Techniques (pp. 196-209), Berlin.

Hicks, T. G., & Wierwille, W. W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors, 21*(2), 129-143.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner Jr, A. C., Zaklade, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors, 34*(4), 429-439.

Hoepf, M., Middendorf, M., Epling, S., & Galster, S. (2015). *Physiological indicators of workload in a remotely piloted aircraft simulation*.

Hunter, J. (2021). *The Truth About The Air Force's Biggest Changes To Pilot Training Since The Dawn Of The Jet Age*. The Drive. Retrieved May 15, 2022 from

Jones, N., Kiely, J., Suraci, B., Collins, D. J., De Lorenzo, D., Pickering, C., & Grimaldi, K. A. (2016). A genetic-based algorithm for personalized resistance training. *Biology of sport, 33*(2), 117.

Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological psychology, 34*(2-3), 237-257.

Kelley, C. R. (1969). What is adaptive training? *Human Factors, 11*(6), 547-556.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International journal of aviation psychology, 3*(3), 203-220.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, 279-328.

Labedan, P., Darodes-De-Tailly, N., Dehais, F., & Peysakhovich, V. (2021, 2021). Virtual Reality for Pilot Training: Study of Cardiac Activity.

Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., & Sarkar, N. (2012). Design of a virtual reality based adaptive response technology for children with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 21*(1), 55-64.

Landsberg, C. R., Mercado, A. D., Van Buskirk, W. L., Lineberry, M., & Steinhauser, N. (2012, 2012). Evaluation of an adaptive training system for submarine periscope operations.

Landsberg, C. R., Van Buskirk, W. L., Astwood Jr, R. S., Mercado, A. D., & Aakre, A. J. (2010). *Adaptive training considerations for use in simulation-based systems*.

Lang, Y., Wei, L., Xu, F., Zhao, Y., & Yu, L.-F. (2018, 2018). Synthesizing personalized training programs for improving driving habits via virtual reality. 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR),

Losey, S. (2021, 23 September 2021). The Air Force Is Still Short on Pilots and Hopes Tech Will Help Close the Gap.

Luo, L., Yin, H., Cai, W., Lees, M., & Zhou, S. (2013). Interactive scenario generation for mission-based virtual training. *Computer Animation and Virtual Worlds, 24*(3-4), 345-354.

Maggio, M. G., Russo, M., Cuzzola, M. F., Destro, M., La Rosa, G., Molonia, F., Bramanti, P., Lombardo, G., De Luca, R., & Calabrò, R. S. (2019). Virtual reality in multiple sclerosis rehabilitation: A review on cognitive and motor outcomes. *Journal of Clinical Neuroscience, 65*, 106-111.

Mariani, A., Pellegrini, E., Enayati, N., Kazanzides, P., Vidotto, M., & De Momi, E. (2018, 2018). Design and evaluation of a performance-based adaptive curriculum for robotic surgical training: A pilot study. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2162-2165),

Marucci, M., Di Flumeri, G., Borghini, G., Sciaraffa, N., Scandola, M., Pavone, E. F., Babiloni, F., Betti, V., & Aricò, P. (2021). The impact of multisensory integration and perceptual load in virtual reality settings on performance, workload and presence. *Scientific Reports, 11*(1), 1-15.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta psychologica, 75*(1), 75-89.

Mazloum, A., Kumashiro, M., Izumi, H., & Higuchi, Y. (2008). Quantitative overload: a source of stress in data-entry VDT work induced by time pressure and work difficulty. *Industrial health, 46*(3), 269-280.

McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016, 2016). Validation of the Empatica E4 wristband.

Metalis, S. A. (1991). Heart period as a useful index of pilot workload in commercial transport aircraft. *The International journal of aviation psychology, 1*(2), 107-116.

Milstein, N., & Gordon, I. (2020). Validating measures of electrodermal activity and heart rate variability derived from the empatica E4 utilized in research settings that involve interactive dyadic states. *Frontiers in Behavioral Neuroscience, 14*.

Monfort, S. S., Sibley, C. M., & Coyne, J. T. (2016). Using machine learning and real-time workload assessment in a high-fidelity UAV simulation environment. Next-Generation Analyst IV (Vol. 9851, p. 98510B). International Society for Optics and Photonics.,

Moray, N. (2013). *Mental workload: Its theory and measurement* (Vol. 8). Springer Science & Business Media.

Moreno, A. (2019). *Combat Jet Cockpit*. https://assetstore.unity.com/packages/3d/vehicles/air/combat-jet-cockpit-74709#description

Morrison, J. E., & Hammon, C. (2000). *On measuring the effectiveness of large-scale training simulations*.

Mühlberger, A., Herrmann, M. J., Wiedemann, G., Ellgring, H., & Pauli, P. (2001). Repeated exposure of flight phobics to flights in virtual reality. *Behaviour research and therapy, 39*(9), 1033-1050.

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology, 34*(2-3), 205-236.

NASA. (2010). NASA Human Integration Design Handbook (HIDH)-NASA (Vol. 3407). *SP-2010*.

Oberhauser, M., & Dreyer, D. (2017). A virtual reality flight simulator for human factors engineering. *Cognition, Technology & Work, 19*(2), 263-277.

Oberhauser, M., Dreyer, D., Braunstingl, R., & Koglbauer, I. (2018). What's real about virtual reality flight simulation? *Aviation Psychology and Applied Human Factors*.

Orlansky, J., Dahlman, C. J., Hammon, C. P., Metzko, J., Taylor, H. L., & Youngblut, C. (1994). The value of simulation for training (IDA Paper P-2982). *Institute for Defense Analyses, Alexandria, Va., USA*.

Pallavicini, F., Argenton, L., Toniazzi, N., Aceti, L., & Mantovani, F. (2016). Virtual reality applications for stress management training in the military. *Aerospace medicine and human performance*, *87*(12), 1021-1030.

Peretz, C., Korczyn, A. D., Shatil, E., Aharonson, V., Birnboim, S., & Giladi, N. (2011). Computer-based, personalized cognitive training versus classical computer games: a randomized double-blind prospective trial of cognitive stimulation. *Neuroepidemiology*, *36*(2), 91-99.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing* Vienna, Austria. https://www.R-project.org/

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, *147*, 103778.

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of experimental psychology: Applied*, *6*(1), 31.

Rehmann, A. J. (1995). *Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research*.

Rogers, R. O., Boquet, A., Howell, C., & DeJohn, C. A. (2007). *An experiment to evaluate transfer of low-cost simulator-based upset-recovery training*.

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological psychology*, *34*(2-3), 259-287.

Roscoe, S. N., & Bergman, C. A. (1980). Chapter 4: Flight performance control. . In *Aviation Psychology*. Iowa State University Press.

Rossol, N., Cheng, I., Bischof, W. F., & Basu, A. (2011, 2011). A framework for adaptive training and games in virtual reality rehabilitation environments.

Saurav, K., Dash, A., Solanki, D., & Lahiri, U. (2018 2018). Design of a VR-based upper limb gross motor and fine motor task platform for post-stroke survivors. 17th International Conference on Computer and Information Science (ICIS) (pp. 252-257),

Schultheis, M. T., & Rizzo, A. A. (2001). The application of virtual reality technology in rehabilitation. *Rehabilitation psychology*, *46*(3), 296.

Schuurmans, A. A. T., de Looff, P., Nijhof, K. S., Rosada, C., Scholte, R. H. J., Popma, A., & Otten, R. (2020). Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: A comparison to electrocardiography (ECG). *Journal of medical systems, 44*(11), 1-11.

Sirevaag, E. J., Kramer, A. F., Reisweber, C. D. W. M., Strayer, D. L., & Grenell, J. F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 36*(9), 1121-1140.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of communication, 42*(4), 73-93.

Summa, S., Basteris, A., Betti, E., & Sanguineti, V. (2015). Adaptive training with full-body movements to reduce bradykinesia in persons with Parkinson's disease: a pilot study. *Journal of neuroengineering and rehabilitation, 12*(1), 1-13.

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

Tattersall, A. J., & Hockey, G. R. J. (1995). Level of operator control and changes in heart rate variability during simulated flight maintenance. *Human Factors, 37*(4), 682-698.

Unity. (2021). *Unity - Manual: Unity User Manual 2020.3 (LTS)*. Unity Technologies. Retrieved 05/15/2021 from https://docs.unity3d.com/Manual/index.html

Veltman, J. A. (2002). A comparative study of psychophysiological reactions during simulator and real flight. *The International journal of aviation psychology, 12*(1), 33-48.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological psychology, 42*(3), 323-342.

Wang, Z., Zheng, L., Lu, Y., & Fu, S. (2016). Physiological indices of pilots' abilities under varying task demands. *Aerospace medicine and human performance, 87*(4), 375-381.

Wilson, G. F. (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological psychology, 34*(2-3), 163-178.

Wilson, G. F., Purvis, B., Skelly, J., Fullenkamp, P., & Davis, I. (1987, 1987). Physiological data used to measure pilot workload in actual flight and simulator conditions.

Wilson, G. F., & Russell, C. A. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors, 45*(4), 635-644.

Wise, J. A., Hopkin, V. D., & Garland, D. J. (2010). *Handbook of aviation human factors* (J. A. Wise, V. D. Hopkin, & D. J. Garland, Eds. 2nd Edition ed.). (2009)

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments*, 27-41.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, *1*(1), 3-28.

Zahabi, M., & Abdul Razak, A. M. (2020). Adaptive virtual reality-based training: a systematic literature review and framework. *Virtual Reality*, *24*(4).

Zhang, J. (2007). Effect of age and sex on heart rate variability in healthy subjects. *Journal of manipulative and physiological therapeutics*, *30*(5), 374-379.

Zhang, Y., & Tsai, S.-B. (2021). Application of Adaptive Virtual Reality with AI-Enabled Techniques in Modern Sports Training. *Mobile Information Systems*, *2021*.

Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surgical endoscopy*, *26*(10), 2746-2750.

Zhou, J., Asteris, P. G., Armaghani, D. J., & Pham, B. T. (2020). Prediction of ground vibration induced by blasting operations through the use of the Bayesian Network and random forest models. *Soil Dynamics and Earthquake Engineering*, *139*, 106390.