Introducing Mental Workload Assessment for the Design of Virtual Reality Training Scenarios

Tiffany Luong, Ferran Argelaguet Sanz, Nicolas Martin, Anatole Lécuyer

Introducing Mental Workload Assessment for the Design of Virtual Reality Training Scenarios

Tiffany Luong*† Ferran Argelaguet† Nicolas Martin* Anatole Lécuyer†

*IRT b<>com, Cesson-Sevigne, France
†Univ Rennes, Inria, CNRS, IRISA, Rennes, France

ABSTRACT

Training is one of the major use cases of Virtual Reality (VR) due to the flexibility and reproducibility of VR simulations. However, the use of the user's cognitive state, and in particular mental workload (MWL), remains largely unexplored in the design of training scenarios. In this paper, we propose to consider MWL for the design of complex training scenarios involving multiple parallel tasks in VR. The proposed approach is based on the assessment of the MWL elicited by each potential task configuration in the training application. Following the assessment, the resulting model is then used to create training scenarios able to modulate the user's MWL over time. This approach is illustrated by a VR flight training simulator based on the Multi-Attribute Task Battery II, which solicits different cognitive resources, able to generate 12 different tasks configurations. A first user study ($N = 38$) was conducted to assess the MWL for each task configuration using self-reports and performance measurements. This assessment was then used to generate three training scenarios in order to induce different levels of MWL over time. A second user study ($N = 14$) confirmed that the proposed approach was able to induce the expected mental workload over time for each training scenario. These results pave the way to further studies exploring how MWL modulation can be used to improve VR training applications.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 INTRODUCTION

Virtual Reality (VR) has known a great breakthrough and is now used in a wide spectrum of applications. It gives the opportunity to fully immerse and engage a user by making them feel "*present*" [53]. VR also allows the design of specific and complex protocols in a safe way which might be not feasible or too expensive in real life. For that reason, it has been used extensively to design training applications. Indeed, it was found that training in immersive Virtual Environments (VEs) could lead to similar performances than training in the real world [62], if not to improve them [2, 5, 27, 42]. From the medical field [1,27,35,52] to the educational field [23,31,38], VR has proven its efficiency as a training tool.

On the other hand, Mental Workload (MWL), which can refer to "*the ratio of demand to allocated resources*" [20], has long been recognized as an important factor within complex systems and in training procedures [12, 19, 25, 39, 41, 43, 66]. In fact, it has been widely used as an offline metric to evaluate the impact of performing tasks to predict operators' performances and system performances [10]. Moreover, it has been supported that optimizing operator's MWL could reduce human errors, improve system safety, increase productivity, and enhance operators' satisfaction [10, 41, 50, 60].

*e-mail: firstname.lastname@b-com.com
†e-mail: firstname.lastname@inria.fr

Several VR studies have explored the modulation of MWL over time by adapting a single task difficulty based on MWL indicators, such as task performance [26, 33, 44] or physiological signals [21]. However, a real training context generally implies complex environments with multiple tasks to perform in parallel. In these cases, the process of adaptation would be more difficult as an increase or a decrease of these indicators might be due to the drop-out or the prioritization of some of the tasks. Thus, identifying which task to adapt to elicit the right amount of MWL is more challenging. For example, aircraft pilots often have to monitor and perform multiple tasks in parallel [51]. They can show overall bad performances but still be underloaded by focusing on only one task. In that case, decreasing the difficulty of the "bad performances" tasks during training might not elicit the expected level of MWL. So far, no strategy was introduced to design VR training scenarios in a multitask context to modulate the users' MWL.

In this work, we propose a new approach to introduce MWL in the design of multitask VR training scenarios. First, it assumes that the training environment is composed of a set of tasks, in which each one can have different levels of difficulty. A task configuration represents the state of all of the tasks at a specific time. All of the task configurations are distinct, and each of them is defined so all of the tasks are represented with one of their difficulty levels. A state machine is created using these task configurations as nodes, and the transitions are designed based on the training scenario constraints. Then, experimentally, the MWL is measured inside each task configuration. Finally, the trainer can define different traversals of the state machine in order to create scenarios which modulate the MWL over time. An application based on a VR flight simulator illustrates the proposed approach. The results showed that this approach was successful into inducing the expected MWL over time for the 3 scenarios designed following the proposed methodology. This approach contributes in making VR trainings more adapted to users' limited cognitive resources [63] by predicting MWL at an early design phase in the conception of VR training scenarios.

The remainder of the paper is structured as follows. Section 2 provides an overview of related work regarding the design of VR training scenarios and MWL in VR training. In section 3, the approach to introduce MWL in the design of training scenarios is presented. An illustrative application using a VR flight simulator is given in section 4. Then, the results are discussed in section 5. Finally, section 6 provides the concluding remarks.

2 RELATED WORK

2.1 Training Scenarios in VR

VE training often implies complex environments with several humanoid agents and interactable objects, which each have a limited number of functions. For that reason, the approaches and models used to design VR scenarios often aim to simplify or to structure the creation of virtual scenarios and mostly are based on narrative or interaction requirements [24]. Those can be divided into two classes: predefined scenario models and emerging scenario models [16].

Predefined scenario models focus on the sequence of the events. They orchestrate the events based on the users' actions and on the characteristics or attributes of the virtual objects. The models rely on diverse representations and are mostly based on automata. Among the different types of representation used in the models of scenarios in VEs, state machines are widely used (e.g., Story Nets [57], HCSM [18]). The events are defined in the states and triggered depending on the users' choices and actions. There are also other types of models which rely, for example, on Petri nets representation such as IVE [7] and #SEVEN [16], graphcet-like representations such as LORA++ [24], or activity diagrams like HAVE [15].

On the other side, **emerging scenario models** do not define precisely what events should occur and in which order. The simulations are driven based on a set of rules that constrains the behaviour of the VE and virtual agents. For example, IDTension [55] and EmoEmma [13] are based on a set of rules which define the actions the agents can undertake (e.g., for IDTension: "wish to realize an objective", "know an information", "can fulfill a task"). Another model, SELDON [11], was thought to adapt dynamically a scenario based on the user's actions. The user can interact freely and the application tries to reorient the scenario toward a specific path by launching events depending on a set of predefined constraints linked to pedagogical and narrative requirements.

In VR training, most studies focused on predefined scenario models as the scenario has already been defined in advance. The use of MWL has been less explored in the design of VR training scenarios.

2.2 Mental Workload Measures

O'Donnell and Eggemeier classified the methods to measure MWL in three groups [43]: subjective (or self-report), physiological, and task performance measures. More details can be found in reviews and surveys on the topic [20, 40, 41].

Self-report methods can be categorized into multidimensional or unidimensional scales. The most commonly used standardized multidimensional scales are the NASA-Task Load Index (TLX) [30], the Subjective Workload Assessment Technique [48], and the Workload Profile [59]. However, those are not adapted to experiments which necessitate several MWL ratings. Among unidimensional scales, the Rating Scale of Mental Effort (RSME) [67] and the Instantaneous Self Assessment (ISA) [56] have been widely validated [10, 20]. The ISA is the quickest to respond as it measures the MWL using five different ratings [56], while the RSME evaluates the mental effort on a continuous vertical axis from 0 to 150 units [67]. Regarding **physiological measures**, a number of studies have shown correlations between MWL and physiological signals such as: electroencephalogram (EEG) [8, 54], electrooculography (EOG), pupillometry, cardiac activity, and electrodermal activity (EDA) signals [4, 61]. As for the **performance measures**, those mainly depend on the type of task. This is also the case for **behavioural measures**, which have shown to be influenced by MWL, especially in VR [8, 9, 37, 49].

2.3 Using Mental Workload in VR Training Scenarios

There are mainly two usages of MWL in studies dealing with VR training scenarios. First, MWL is majorly used as an offline metric in VR studies. Users are exposed to a VR training scenario once or multiple times and their performances and MWL are being assessed [6, 36]. The purpose can be to show VR training can be as efficient or more efficient than more traditional methods [2, 5, 27, 42, 62] or to study the impact of other independent variables on the MWL [36]. Usually, the training scenarios are designed specifically for the study objective and can not be easily re-adapted without further coding and research.

On the other hand, MWL can be used to adapt the difficulty in the training scenario. For example, task performances can be used to assess MWL. Grimm et al. [26] adapted the level of difficulty of a reach-to-grasp task based on the performance of chronic stroke patients by adapting the distance between the object to grab in VR and the target where the patient had to release it. In the same line, Kizony et al. developed 4 different clinical applications in VR, each featuring one task which difficulty could be adapted based on the user's performance [33]. In the military field, Parsons et al. proposed a framework, which adapted the complexity of a Humvee following task by varying the vehicle's acceleration and deceleration based on the distance between the user's vehicle and the following vehicle [44]. Alternatively, physiological signals have been found to be highly correlated to affective and cognitive states [22, 45], in particular, mental workload [29, 34, 61]. This has been of great interest since decades outside the VR community, and some VR studies focused on the classification of affective or cognitive states based on physiological signals such as fNIRS [47], EEG [58, 65], EDA and heart rate signals [17, 65]. As such, in the context of difficulty adaptation, Dey et al. adapted the difficulty of a visual searching task based on the power of the alpha band of an EEG [21]. However, it should be noted that while performance, physiological, and behavioural measures are continuous and do not require a formulated response by the subjects, those tend to lack of accuracy and are difficult to generalize [43, 60]. On the other hand, self-report methods enable high-face validity and are a direct measure of the user's cognitive state [10, 20], which makes them preferable for the design of scenarios based on MWL.

While there exist models which facilitate the creation of training scenarios as sequences of events, they all rely more on constraints linked to the narrative requirements and to the interactions between the virtual objects and agents than on the users' cognitive state. No approach nor model has been proposed to generate scenarios involving multiple tasks which modulate the MWL over time for immersive VEs to our knowledge yet. Moreover, the generation of scenarios based on the progression of MWL in a complex context with multiple parallel tasks could benefit trainers, who do not always have the time nor the resources to customize VR training applications for new users or new training objectives.

3 Methodological Approach

In this section, we propose an approach to introduce MWL in the design of VR training scenarios to make trainings more fitted to users' cognitive state [12]. The goal is to modulate the user's MWL over time in complex contexts involving multiple parallel tasks.

Our approach is based on the model of the states of the training application using different task configurations. State machines are convenient in our proposed approach as it can be used to generate scenarios based on the arrangement of tasks configurations (i.e., nodes or states) over time. Each state contains data about the tasks setting and the users' mental workload, and the transitions help to constrain the sequencing of the tasks.

This methodology is divided into the following steps. First, tasks are selected based on the objective and on the context of the training study. Then, each selected task is subdivided into different difficulty levels. These task levels are combined to form the states of a state machine. The state machine transitions are defined to allow users to navigate from a task configuration (i.e., a state) to another. Following these steps, the experimenter chooses which MWL data to collect. The MWL is thereafter assessed and will determine the attributes of the states. These measures can finally be used to design training scenarios by defining paths inside the state machine.

3.1 Tasks Identification and State Machine Generation

The objective of the first steps of the methodology is to design the states and the transitions of the state machine. The states should contain data about the tasks configuration and the MWL they induce to users. The transitions (i.e., a vector with a start and an end states)

should determine if users can transit from one state to another or not.

Starting with the tasks selection, those should be defined based on the training context. For example, in a car manufacturing training scenario, the main task can be the assembly of car pieces and in a fire safety application, the main task can be the extinction of a fire situation. The idea of this first step is to extract all tasks relevant to a training purpose in the imposed context. Once the tasks are selected, those can be decomposed into multiple discrete intrinsic levels of difficulty, including their presence (activation) or absence (deactivation), if relevant. Each task level should be combined to form the "states", except for the tasks levels which can not be associated given the training environment. In any case, all tasks should be represented in each state with a specific level of difficulty.

In the same way, the transitions should be created depending on the constraints of the training context. By default, those can be defined so only one task at a time can be upgraded or downgraded in its difficulty level, or so all states are linked together. However, if one task level can only come after another specific task level, this should be taken into account. The condition to trigger a transition can be of multiple types (e.g., a time limit, once all the tasks of a state have been fulfilled, after users have reported their MWL...).

3.2 Data Assessment

Following the identification of the tasks levels, the trainer has to choose which MWL data to collect (see Section. 2.2). Then, the objective is to assess a maximum of data to have an estimation of the MWL induced by each state.

For a clean estimation and to avoid order effects, the experimenter can make all participants traverse the states in a randomized or counterbalanced order so each state is explored in a balanced way. In the case the training context would be too complex and the states too numerous, only certain states can be used to assess data. For example, for tasks with more than 2 levels of difficulty, the intermediary states considering a variation of difficulty of one task can be ignored during the assessment and estimated afterwards. Nevertheless, this will draw to less accuracy about the MWL induced by these states.

Once the data is measured, it can be assigned to the states as attributes. At this point, the states should contain data about the task configuration (i.e., *"task1 difficulty level"*, *"task2 difficulty level"*, ...) and about the mean mental workload induced by the task configuration to users (i.e., *"MWL measure 1"*, *"MWL measure 2"*, ...). The measured data can be used to weight the transitions (e.g., the difference between the subjective mental workload of the end state and of the start state) for the scenarios design purposes. Also, a confidence value can be attributed based on the MWL data distribution inside a state.

3.3 Scenario Generation

The MWL measures can now be used to design scenarios by defining the order of the task configurations (i.e., states) through time.

From this point, the keys of the design belong to the trainer. Depending on the motive of the training, they have to define the scenario in function of the MWL they want to induce to the trainee. The trainer can as well make a list of constraints linked to the duration of the training, the number of simultaneous changes of tasks for the transitions, and the appearance or not of a task level. The computation of the paths can be done using the transitions weights (which can be defined as the difference of MWL data between the end and the start state). For example, if a scenario should maintain the user's mental workload at the same level all long, the path can be computed by choosing the transitions with the cheapest costs (i.e., lowest differences of MWL).

This last step should result in the design of a training scenario which modulates the progress of MWL over time, as defined by the trainer.

4 Illustrative Application: Training Scenarios Based On MWL in a VR Flight Simulator

To illustrate the proposed approach, we designed a VR application based on a flight simulator. Two experiments ($N1 = 38$, $N2 = 14$) were executed for this purpose. The objective of the first experiment was to collect MWL data in the defined task configurations. The objectives of the second experiment were to use these measures to create scenarios based on training objectives, and to compare the MWL results to the expected outcomes.

This section is structured following the proposed methodology in section 3. First, tasks and their difficulty levels are defined following the context of the training. Those are combined to form states, which are used to structure a state machine. Then, the MWL measures are identified and assessed through a first experiment. Finally, these data and the given states are used to design 3 scenarios.

4.1 Tasks Design

The illustrative application was designed based on a VR flight simulator. It is inspired by the second version of the Multi-Attribute Task Battery (MATB-II) [51], a computer-based application designed to induce and evaluate an operator's performance and workload developed by NASA. This battery of tasks has been widely used to study multitasking and the use of automation [51]. The original application comprises a monitoring task, a tracking task, a schedule window, a communication task, and a resources management task. Those were intended to be presented on a single computer window and are analogous to tasks that aircraft crew-members perform in flight, while being accessible to non-pilot subjects.

In the current studies, three tasks were selected and adapted to a VR environment: the piloting task, which would be an analogy to the tracking task, the communication task, and the resources management task (see Fig. 1).

Figure 1: Virtual cockpit view. (1) Instantaneous Self Assessment (ISA) interface. (2) Resources management task interface (deactivated); when activated, the interface lit up (with a green outline). (3) Communication task interface (activated); when deactivated, the interface lit off (no green outline). (4) Informative panel which gives information about which task is activated or not at the current time. (5) Virtual representation of the joystick used to pilot the aircraft and of the right hand. The left hand is represented in the same way, but tracked by a Vive Controller and animated depending on the interaction.

Piloting Task

The tracking task of the MATB-II required the user to keep a target in the centre of a square controlled by a joystick. In the following experiments, the task was adapted to fully exploit the immersion permitted by VR. Participants were piloting an aircraft in the first-person perspective using a joystick and could see the

environment being refreshed in real-time depending on their actions. The speed of the piloting task was imposed on the users (they could not accelerate nor decelerate). They could, however, orientate the aircraft using a joystick. Preliminary tests were done to tweak the sensitivity and the control of the interaction. Since the participants were non-pilot subjects, only two degrees of freedom of the aircraft were retained: the yaw (i.e., rotation upon the vertical axis) and the pitch (i.e., rotation to go up and down). The roll degree (i.e., rotation upon the forward vector of the aircraft) was not included. The objective of this task was to follow as closely as possible the green line which passed through all circles centres.

Three levels of difficulty were proposed: easy piloting (0), medium piloting (1), and hard piloting (2). The task difficulty was modified by adapting the aircraft speed and the number of visible circles at a time. In the easy difficulty, users could see 3 circles at a time and the aircraft was advancing at a slow pace (about 20 seconds between 2 circles). In the medium difficulty, this was changed to 2 circles and a speed which was multiplied by about 2, and in the hard difficulty, to one circle and an original speed which was multiplied by about 3.

Each circle was separated by a constant distance. Only the easiest task configuration was set to have a straight alignment of circles. Otherwise, the trajectory was randomized for each participant so the horizontal and vertical distances between two circles could not exceed an imposed value.

Communication Task

The communication task was designed similarly to the one from the MATB-II, but without the radio channel selection.

This task was designed to have two levels: it was either disabled (0) or enabled (1). When the task was enabled, audio messages were sent to the participant's audio headset, and the interface of interaction lit up (see Fig. 1).

The operators were asked to answer when the message was intended for their aircraft. In the case of our experiment, the ID of the operator's aircraft was "HDG219". When the message was directed to the aircraft, the participant was required to change the frequency of the radio in accordance with the message, by pushing the "plus" and "minus" buttons, which changed the left screen (see Fig. 1). Once users finished inputting the frequency, they were asked to click on the validate button, which changed the right screen so it matched the left screen (see Fig. 1). No action was expected when the message was directed to another aircraft. In the studies, half of the calls in each state with the communication task were set to target the user.

The target frequency was an integer number of 3 digits with no decimal part. It was computed so the user was asked to click on the plus or minus button (see Fig. 1) a random number of times between 2 and 5 included. When the message was not directed to the user, the requested aircraft was a random one between 4 different IDs.

Resources Management Task

This task was designed in accordance with the one from the MATB-II. It depicts a generalized fuel management system.

This task had two levels: disabled (0) and enabled (1). When the task was enabled, the interface of interaction of the task was lit up (see Fig. 2 and Fig. 1).

The interface displayed 8 different pumps numbered from 1 to 8, and 6 tanks labelled from A to F (see Fig. 2). When the task was enabled, the fuel of the tanks A and B started to decrease. The objective of the user was to keep the levels of the tanks A and B in the blue zone displayed on the two tanks sides. Those indicated the critical levels of fuel for those tanks. Tanks D and F had unlimited capacities. Not transferring fuel to tanks A and B resulted in empty tanks after some time, while tanks C and D only lost fuel if they were transferring fuel to another tank.

Users could use the 8 pumps at their disposal to maintain the two tanks levels in the appropriate zone. There were 3 possible states for each pump button: grey, which meant the pump was deactivated, green, which meant the pump was activated and red, which meant the pump was failed and could not be used. An activated pump button (green) meant fluid was circulating in the direction indicated by the button arrow. Clicking on a grey button would turn the button green, and clicking on it again would turn the button back in grey, like a switch button. The red state was activated at predefined moments.

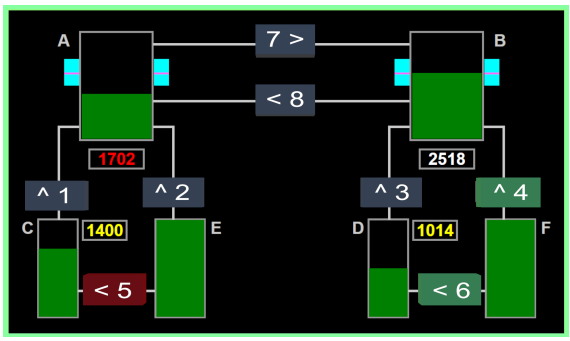

Figure 2: Resources management task interface. A and B are the main tanks; their fuel levels are indicated below the tanks. C and D are supply tanks; their fuel levels are indicated on their right side. E and F are supply tanks with unlimited capacities. The buttons numbered from 1 to 8 are pumps button. Pumps 4 and 6 are activated, pump 5 is failed, and all other grey pump buttons are deactivated.

4.2 State Machine Structure

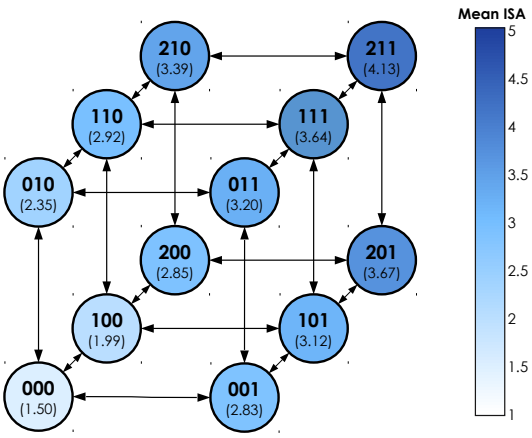

Figure 3: State machine of the designed VR flight simulator. Each state is labelled using 3 digits. The first digit refers to the level of the piloting task (0-easy, 1-medium, 2-difficult), the second one to the level of the communication task (0-deactivated, 1-activated) and the third one to the level of the resources management task (0-deactivated, 1-activated). The colours of the nodes represent the mean ISA assessed during the first study, which are as well indicated in the brackets. Only the transitions between consecutive states are depicted there.

As seen in Section 4.1, 3 tasks were considered, each of them with their own difficulty levels. This gives us: 3 piloting task (0-easy, 1-medium, 3-hard) $\times$ 2 radio task (0-deactivated, 1-activated) $\times$ 2 resources management task (0-deactivated, 1-activated). All these levels were combined to form 12 states ($3 \times 2 \times 2$). To ease the understanding, each state was labelled using 3 digits, one for each task. The first digit represents the difficulty of the piloting task,

the second one, the level of the communication task, and the third one, the level of the resources management task. For example, in the state "201", the difficulty of the piloting task is set to "2-hard", the radio task is "0-disabled" and the resources management task is "1-enabled". Following this labelling system, the resulting state machine is depicted in Figure 3. For convenience purposes, we will call "consecutive states" two states which differ in only one task of one difficulty level. All the states could transit from one to another (i.e., there were transitions between all states). The transitions were triggered once users passed the last circle of the state they were in.

4.3 Mental Workload Measures

The dependent variables considered were: self-reports (i.e., the subjective MWL reported by the participants) and task performances.

Self-Reports: Originally, the MATB-II was designed so the MWL could be assessed using the NASA-TLX [30] after the experiment. However, the latter is a multidimensional scale, which is not adapted to the several ratings of the MWL throughout a scenario. Therefore, the focus was set on the ISA [56], which rates the MWL using five different ratings (under-utilized, relaxed, comfortable, high, excessive) and has been especially used during air traffic control tasks [56]. The meaning of each rating levels of the ISA was explicited to each subject using the description given by Tattersall et al. [56] before each experiment.

During the experiments, users were asked to report their MWL when a screen appeared in front of them with 5 buttons (see Fig. 1). They were asked to push the button corresponding to their MWL level and then, to click on the validate button on the same screen to make it disappear.

Performance Measures: Concerning the piloting task, the distance to each circle centre was recorded throughout the experiment. The participants were given indication before the experiment of how to align the plane with the green line optimally.

As for the communication task, performance indicators of the success of the task were recorded: true positive and negative, accuracy, difference with the target frequency, reaction and response time, global success for each communication call compared to the expected reaction.

Lastly, the global success time ratio of the resources management task was recorded as well as events corresponding to a success or a fail of the task.

4.4 Apparatus

Each participant was installed on a cockpit assembled for the experiment (see Fig. 4). The virtual environment was modelled in 3D so the virtual cockpit matched the real one in position and size.

As for the interactions, users piloted the aircraft using a Logitech 3M X52 PRO with the right hand, and they interacted with the virtual interfaces using a Vive controller with the left hand. Both hands were represented by transparent virtual hands. The right hand was placed on the virtual joystick (see Fig. 1), which moved when the user was interacting with the real one. The left hand was tracked with the Vive controller. As seen in Section 4.1, all interactions linked to the tasks are buttons based. The users did not need to use any of the buttons on the Vive Controller or on the joystick to do the tasks. Pushing a virtual button would be processed as follows: the user approaches her.his left hand to the button. The interactable object highlights and the hand animation changes to a pointing index. The user pushes further the button as s.he would have done with her.his real finger. This triggers a small haptic pulse on the Vive Controller, which gives a feedback that the action has been well resolved. The environment updates according to the button action.

The VR headset used during the experiments was a Vive Pro. Audio instructions were provided using the audio headset supplied with the HMD. The support application was developed in Unity 3D, and run on a laptop computer equipped with an Intel(R) Core(TM) i7-6820HK CPU (2.7 GHz), one Nvidia GeForce GTX 1070 graphic card, and 16 Go Random-Access Memory.

Figure 4: Picture of the experimental set-up. The user is wearing a Vive Pro, and using a joystick and a Vive controller on the cockpit.

4.5 Experiment 1: Data Assessment

The objective of this first experiment was to collect data on the users' MWL in each state (i.e., tasks levels combination) to get an overview of their effect on the MWL (see Fig. 3).

4.5.1 Participants

39 healthy participants from our research institute volunteered to take part in this study. One subject was excluded from the study because of motion sickness issues, resulting in a final sample of 38 participants (10 females, 28 males; ages 21-62, M=36.97). Four participants reported having a small experience with aircraft piloting, and all others never had any experience with it. All participants were fluent in French and were naive to the experiment conditions and purpose. They all completed and signed an informed consent form before the start of the experiment.

4.5.2 Experimental Design

The goal of this experiment was to measure the users' MWL induced by each state of the graph. Thus, each user went by all the states. Only transitions between consecutive states were considered in this study. To minimize the learning effect, the order of the states was randomized for each participant. However, they always started with the state "000" (easy piloting, no communication task, no resources management task).

The objective was to make sure that the designed states could induce different levels of MWL. Therefore, we hypothesized that:

- **H1**: Increasing the level of difficulty for each task will increase the users' MWL.
- **H2**: Additional task will increase the users' MWL.
- **H3**: The subjective MWL will have an effect on the tasks performances.

4.5.3 Experimental Procedure

The experiment had a total duration of around 1h and was subdivided into the following steps:

Written Consent and Instructions: Users completed a consent form, prior to the experiment. They were then instructed with the nature of the experiment, the equipment used, the data recorded (which was anonymized), and the tasks instructions. Participants were also asked to fill a questionnaire (experience with VR, video

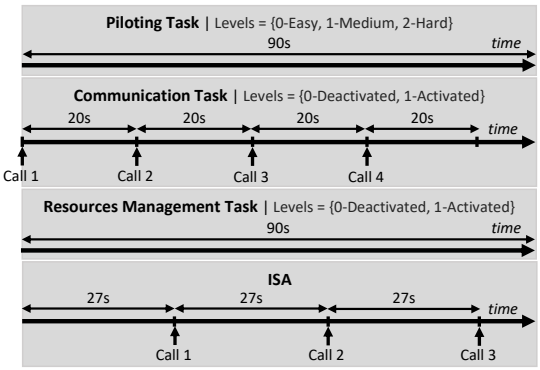

Figure 5: Progress of the tasks and of the ISA calls within one state in the first experiment. In this paper, 12 states were considered. The communication task and the resource management task were not always activated. The ISA is always present. Each state lasted at least 90 seconds (depending on the user's piloting path).

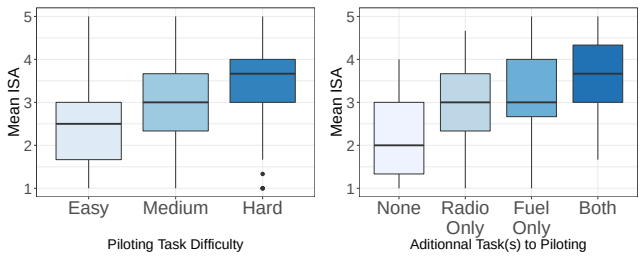

Figure 6: Mean ISA depending on the tasks level and configurations. "Radio" refers to the communication task and "Fuel", to the resources management task.

games, and piloting an aircraft, dominant hand, level of alertness, state of vision, demographic information, simulator sickness questionnaire (SSQ) [32]) to gather information about their background and their state before the start of the experiment.

Training: Users were then equipped with a Vive Controller and the HMD. They were first asked to interact with the buttons of the tasks interfaces to familiarized themselves with the interactions. Then, they travelled the states following this path: "$000-010-011-001-101-201-211$" (see Fig. 3), which gave them a good overview of each task and their levels.

The training part was followed by a 2-minutes pause where users did not wear the HMD and were invited to ask any question they may have had.

Experiment: In this experiment part, users travelled all the states in a randomized order, starting with the state "000". The states were set to last 90 seconds with 4 communication calls and 3 ISA calls. The progress of the tasks within a state is depicted in Fig. 5.

Debriefing: At the end of the experiment, they were asked to fill the SSQ again, debriefed and invited to ask questions.

4.5.4 Results and Discussion

Generalized linear mixed model (GLMM) analysis was considered for all dependent variables (all parametric). For each variable, the user was considered as a random factor and all the independent variables as within-subject factors. When the equal variances assumption was violated, the degrees of freedom were corrected using the Greenhouse-Geisser method. When needed, pairwise post-hoc tests (Bonferroni with adjustment) were performed. Only significant differences ($p < 0.05$) are discussed. The statistical analysis was performed using the R statistical software.

ISA: The GLMM showed a main effect of the piloting difficulty $F_{1.83,67.62} = 100.77$, $p < 0.001$, $\eta_p^2 = 0.73$, communication task $F_{1,37} = 85.23$, $p < 0.001$, $\eta_p^2 = 0.70$, and resources management task $F_{1,37} = 154.68$, $p < 0.001$, $\eta_p^2 = 0.81$. It also showed an interaction effect between the communication task and the resources management (fuel) task $F_{1,37} = 12.81$, $p = 0.001$, $\eta_p^2 = 0.26$, as well as an interaction effect between the piloting difficulty and the fuel task $F_{1.81,67.39} = 3.72$, $p < 0.05$, $\eta_p^2 = 0.09$. Post-hoc tests showed that as the level of difficulty increased, participants perceived the task as more mentally demanding (see Fig. 6) (all $p < 0.001$). The tasks were also perceived as more demanding when a task was activated (see Fig. 6). The resources management task was perceived as more difficult than the communication task ($p < 0.01$; all $p < 0.001$ for other tasks combinations effects otherwise ; see Fig. 6). These results support **H1** and **H2**.

Piloting task performance: The considered variable is the distance to circles centres. The GLMM showed a main effect of the piloting difficulty $F_{1.06,39.10} = 16.07$, $p < 0.001$, $\eta_p^2 = 0.20$, communication task $F_{1,37} = 16.33$, $p < 0.001$, $\eta_p^2 = 0.31$, and resources management task $F_{1,37} = 19.95$, $p < 0.001$, $\eta_p^2 = 0.35$. It also showed a main effect of the mean ISA value (mean of the 3 ISA values per state) $F_{1,447.80} = 46.30$, $p < 0.001$, $\eta_p^2 = 0.08$. Post-hoc tests showed that as the difficulty level increased, the participants tended to go further away from the circles centres (all $p < 0.01$ except between easy and medium difficulties). In the same way, the activation of the communication task or of the resources management task increased the distance to the circles centres. Overall, the more the MWL increased (ISA value), the greater the distance to the circles centres was.

Resources management task performance: The considered variable is the success time ratio of the task. The GLMM showed a main effect of the activation of the communication task $F_{1,37} = 7.52$, $p < 0.01$, $\eta_p^2 = 0.17$, and of the mean ISA value (mean of the 3 ISA values per state) $F_{1,219.64} = 31.15$, $p < 0.001$, $\eta_p^2 = 0.10$. Overall, the activation of the communication task decreased the success time ratio of the resource management task. Moreover, as the MWL increased, the success time ratio of the resources management task decreased.

No significant effect was found on the communication task performance indicators.

Discussion: The levels of the tasks were able to induce different levels of MWL (see Fig. 6). Fig. 3 depict a global map of the mean MWL assessed during this first experiment. Both **H1** and **H2** were supported in our experiment. The resources management task induced a higher MWL than the communication task. This effect could be explained by the fact it stimulated similar pools of cognitive resources to the ones stimulated by the piloting task (visuo-motor) compared to the ones stimulated by the communication task (mainly auditory-motor; see Wicken's Multiple Resources Theory [63]).

The subjective results were supported by piloting and resources management tasks performances, which supports **H3**. However, those were less steady in the states compared to the self-report measures. No significant effect was found for the communication task as it was mainly used to distract the user from the other tasks and to induce time load. Participants managed to successfully complete the task most of the time.

Overall, the designed tasks configurations were successful into inducing different levels of MWL to the users. Moreover, the subjective MWL and the states had an effect on the piloting task performance (distance to the circles centres) and on the resources management task performance (success time ratio). However, the ISA value was more reliable to differentiate the different levels of mental workload in the states than the performance measures considering the effect of the states and the data distribution. Finally, each state was characterized by the following attributes: *(piloting task level, communication task level, resources management task level)*; *(ISA value, piloting task performance, resource management task performance)*. As **H3** was partially supported, we only used the *ISA value* among the MWL measures in the subsequent study.

4.6 Experiment 2: Scenarios Generation

The objective of this second experiment was to build scenarios based on the states defined in Section 4.5 and to validate the proposed approach.

4.6.1 Participants

14 healthy people from our research institute, who did not participate in the first experiment, volunteered to take part in this study (2 females, 12 males; ages 21-52, M=32.21). None of the participants had an experience with piloting an aircraft in the past. As in the first experiment, all participants were fluent in French and were naive to the experiment conditions and purpose. They all completed and signed an informed consent form before the start of the experiment.

4.6.2 Experimental Design

We chose to focus on 3 different scenarios using a fixed number of 5 states. They were all designed based on the subjective MWL (ISA).

The **scenario 1** *induces a medium MWL level all long.* This scenario is a classical use case where a trainer wants to keep the users' MWL at a medium level during the whole training, while varying the experimental conditions. The goal can be to train users while keeping them engaged, by not overloading nor underloading them, and without using repetitive tasks. This scenario was designed by choosing the 5 states with the MWL the closest to $ISA = 3$. The transitions between each of these states were computed so there was a prioritization of transitions between consecutive nodes. The orientation of the given path was chosen randomly. The resulting path is: 200 - 101 - 001 - 011 - 110 (see Fig. 3).

The **scenario 2** *induces a low MWL level first, and a suddenly high MWL level in the end.* This scenario is a typical scenario where the reactivity of the pilot is studied in an urgency or surprising situation. The 4 first states were chosen as the 4 consecutive states with the lowest ISA. The first state was chosen by prioritizing the state "000", and the orientation of the path was chosen randomly. The last state was chosen as the one with the highest ISA value. The resulting sequence is: 000 - 010 - 110 - 100 - 211.

The **scenario 3** *induces a progressively increasing MWL level throughout the experiment.* It can be used to train progressively pilots to different levels of difficulty they might experience. The scenario was defined so the first state was the one with the lowest ISA, and the last state, the one with the highest ISA. All paths of 5 consecutive states going from the first state to the last one where computed, as well as the sum of the transitions weights for each path. The final path was selected as the one the lowest sum of transitions weights: 000 - 100 - 110 - 210 - 211.

This experiment followed a one-factor (scenario) within-subject design. All participants experienced all 3 scenarios. The order of the scenarios was counterbalanced using a Latin square design, except for the 2 last participants who did the 3 scenarios in a random order.

Concerning the hypotheses, they are defined by the scenarios depicted above. We expect no significant effect of the experiment (first or second one) on the mean ISA of the states.

4.6.3 Experimental Procedure

The procedure is almost the same as the one defined in Section 4.5.3, except the SSQ was not included this time. All participants were however asked how they felt after the scenarios.

The experiment was divided into 3 blocks: one for each scenario. After each scenario, users were asked to answer a set of custom questions on a 7-points Likert-Scale (1: fully disagree, 7: fully agree) to define their perceived difficulty during the experiment: *(Q1) I felt that the difficulty was approximately the same during the scenario*, *(Q2) I felt the difficulty suddenly increased at a given time*, *(Q3) I felt the difficulty increased progressively*. These questions respectively transcribe the progress of the scenarios 1, 2, and 3, from MWL to perceived difficulty. They were mixed with 3 other unrelated questions linked to the users' absorption, and sense of presence [53] to mask the interest of the study to participants. At the end of the 3 scenarios, participants were asked to map the descriptions of the 3 scenarios to the 3 experiences they did.

The parameters used in this experiment were set so each state lasted 70 seconds, with 3 communications calls (2 which targeted the user) and 2 ISA calls.

4.6.4 Results and Discussion

ANOVAs were run to analyse the results. To ensure the replication of the first experiment results, the outcomes between experiment 1 and 2 (between-subject factor "Experiment") were compared. For each scenario, no significant effect was found on the ISA (see Fig. 7) and performances. As well, each state in each scenario was also tested separately considering the experiment factor, and it did not show any significant effect on the ISA and performances.

The subjective perceived difficulty after each scenario was as well assessed. For the first scenario, the mean ratings were $\mathbf{(Q1 : 3.36)}$, $(Q2 : 3.14)$, and $(Q3 : 2.64)$. For the second scenario, the results gave $(Q1 : 2.43)$, $\mathbf{(Q2 : 5.86)}$, and $(Q3 : 4.14)$. As for the third scenario, the average scores were $(Q1 : 1.64)$, $(Q2 : 3.79)$, and $\mathbf{(Q3 : 5.71)}$. For each scenario, the related question was noted the best among the 3 in average. The second and the third scenario were globally successfully noted based on the MWL they were supposed to induce, but the first scenario did not induce the right subjective perceived progress of difficulty. As for the mapping of the scenarios to the appropriate experiment, the first scenario was successfully mapped at 85.71%, the second one at 71.43%, and the third one a 64.29%. A confusion was observed mainly between the second and the third scenarios.

The ISA scores and the task performances followed the expected results for scenarios 1, 2, and 3, which were designed based on the first experiment measures (see Fig. 7). The slight variations between the two experiments can be explained by the fact there might have been a small order effect as the travel of the states was randomized in the first experiment and not in the second one. However, as tested, those are non significant. As well, the subjective measures were once again supported by the piloting and the resources management tasks performances.

The perceived difficulty was found to lack of accuracy compared to the expected outcomes, on the contrary to the ISA values. This can be due to the fact these questions were non standardized and to the difference in the rating delays [64]. Some users were found to answer in unexpected ways compared to their ISA score, even just after having completed the scenario. As well, the fact there was a confusion between the mapping of the second and the third scenario at the end of the experiment might be explained by the fact users mainly remembered the first and the last states of the scenarios, which appear to be the same in the two scenarios. Therefore, subjective ratings have to be performed cautiously and delays in rating, considered carefully [64]. The subjective MWL rating throughout the scenarios was more accurate than the post-experiment self-report questions and the performance measures.

5 General Discussion

In this paper, we proposed an approach to create scenarios in order to induce different levels of MWL during a virtual reality training routine. It differs from the literature as it focuses directly on the user's cognitive state and not on narrative or interaction requirements, nor on performance measures. The methodology is divided into the following parts: the tasks levels identification and association, the MWL assessment, and the scenarios design. Two studies based on a VR flight simulator were performed to test the method.

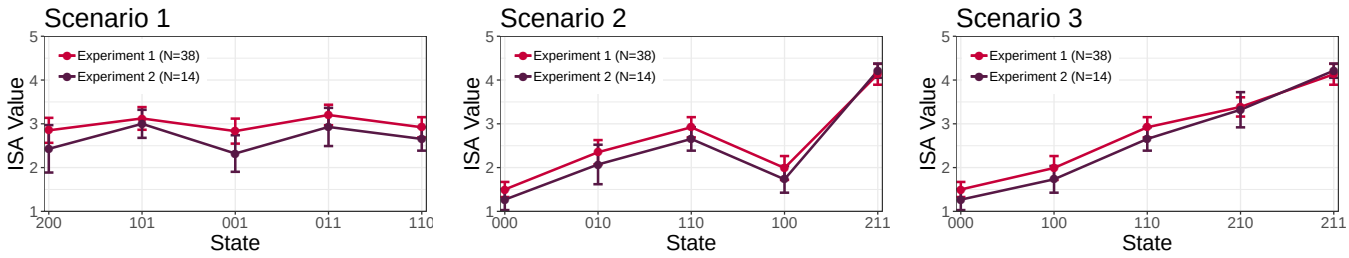

Figure 7: Comparison of the users' subjective MWL (mean ISA value) for each 3 scenarios between the first and the second experiment. The first scenario was intended to induce a MWL around $ISA = 3$ all long and the second one, to induce a low MWL ($ISA <= 3$) followed by a high MWL at last. The last scenario objective was to induce a progressively increasing MWL.

The first study results support that the designed states (i.e., tasks configurations), which are based on the combination of multiple tasks with varying levels of difficulty, were able to induce different levels of MWL. This is consistent with previous findings [14, 28, 36, 63]. They also support that using two tasks stimulating the same pools of cognitive resources will further increase MWL compared to two tasks stimulating different pools of cognitive resources, which goes in the same line as Wicken's Multiple Resources Theory [63]. Both subjective and task performance measures were influenced by the states, but the subjective MWL measures appeared to be more accurate, as outlined in [10, 12, 66]. From the subjective MWL measures gathered and the designed states, 3 training scenarios inducing different progressions of MWL through time were generated in a second study. Those were able to reproduce the same MWL profiles than in the first experiment. However, some users did not perceive the entire scenario difficulty progression after the experiment as accurately as their subjective MWL was predicted over time, which can be explained by the difference of delays in ratings [64]. Overall, the approach was successful into designing training scenarios which were able to induce the expected MWL through time. This supports the use of MWL in training scenarios, as expressed by other studies [12, 14, 36], and encourages the use of MWL in the design of VR training scenarios to make them more fitted to users' cognitive resources in a controlled, reproducible, and safe environment. On a side note, in the two presented studies, most users reported having enjoyed the training and being engaged throughout the experiment, which encourages the use of VR in the training field and is in line with previous studies [2, 3, 5, 14, 27, 42, 46, 62].

In the second study, the 3 scenarios were designed based on the mean subjective MWL assessed in the first experiment and considering realistic training purposes. However, the scenarios could also have been constrained by other criteria such as task performance measures, the presence and the absence of one task, or the probability to induce a subjective MWL rating level for example. Given the designed states with just subjective and task performance measures, there are already numerous ways to generate scenarios. As such, if the first scenario ($ISA \approx 3$ during the whole scenario) appears to be too easy or too difficult for some outliers trainees, a similar design process can easily be replicated to generate a new scenario which induces a higher or a lower MWL all long. Also, we can fully imagine combining our approach to other tools developed to simplify the design of scenarios in VEs which focus more on narrative and interaction requirements, by introducing more complex mechanisms, labels and constraints for example (see Section 2.1).

Concerning the limits of the approach, the assessment part can be time-consuming depending on the complexity and the number of tasks, even if it helps accelerating the design of a wide variety of training scenarios afterward. Some simplification can be made by assessing the MWL only in strategical nodes, and by inferring the others (see Section 3.2). However, this will draw to less accuracy in the prediction of MWL throughout the scenarios. Also, in the first study, the transitions were only set between consecutive nodes, which was not the case in the second study. It did not have an effect on the MWL measures. Yet, the transition effects might have to be taken into account in some contexts. Finally, the self-report request (ISA, see Section 4.3) could be considered as a fourth task as users were required to answer each time they saw the screen appearing. This probably increased their MWL in a similar way for each state as tests did not show an effect of the tasks levels on the ISA response time.

Future works could be dedicated to the use of this approach to understand the effect of MWL modulation on users in VR training. Also, only discrete task difficulty levels were considered here. Supplementary work could be done to investigate the possibility to extend the approach to continuous task difficulties. Furthermore, it would be interesting to extend the approach to the adaptation of scenario in real-time based on the MWL. In addition, this paper focused on MWL, but a similar approach could be imagined for other cognitive states, fatigue, stress, or affective states in other contexts, to build scenarios based on the evolution of users' psychological or physiological state over time.

6 Conclusion

This paper proposes an approach to introduce mental workload measures in the design of complex training scenarios involving multitasking in VR. First, tasks levels are identified and associated. Then, mental workload is assessed inside each task configuration. Finally, the training scenarios can be designed based on the mental workload measures. This approach allows the generation of different training scenarios based on the progression of mental workload over time using different task levels combinations.

Two studies based on a VR flight simulator were performed to test the approach. Taken together, the results support the approach was successful into designing training scenarios which induced the expected progression of mental workload through time. These results pave the way to further studies exploring how mental workload modulation can be used to improve VR training applications.

Acknowledgments

Our study was carried out within b<>com, an institute of research and technology dedicated to digital technologies. It received support from the Future Investments program of the French National Research Agency (grant no. ANR-07-A0-AIRT). We would like to thank Mathieu Quentel, Océane Cloarec, Pierre Le Gargasson, and Grégory Hocquet for their technical support, as well as Flavien Lécuyer for his feedbacks.

References

[1] G. Ahlberg, L. Enochsson, A. G. Gallagher, L. Hedman, C. Hogman, D. A. McClusky III, S. Ramel, C. D. Smith, and D. Arvidsson. Proficiency-based virtual reality training significantly reduces the error

rate for residents during their first 10 laparoscopic cholecystectomies. *The American journal of surgery*, 193(6):797–804, 2007.

[2] W. Barfield, C. Hendrix, and K.-E. Bystrom. Effects of stereopsis and head tracking on performance using desktop virtual environment displays. *Presence: Teleoperators & Virtual Environments*, 8(2):237–240, 1999.

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

[4] W. Boucsein. Electrodermal indices of emotion and stress, chapter 3. *Electrodermal Activity*, pp. 369–391, 1992.

[5] D. A. Bowman and R. P. McMahan. Virtual reality: how much immersion is enough? *Computer*, 40(7):36–43, 2007.

[6] J. Bric, M. Connolly, A. Kastenmeier, M. Goldblatt, and J. C. Gould. Proficiency training on a virtual reality robotic surgical skills curriculum. *Surgical endoscopy*, 28(12):3343–3348, 2014.

[7] C. Brom and A. Abonyi. Petri-nets for game plot. In *Proceedings of AISB artificial intelligence and simulation behaviour convention, Bristol*, vol. 3, pp. 6–13, 2006.

[8] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld. Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[9] G. Bruder, P. Lubas, and F. Steinicke. Cognitive resource demands of redirected walking. *IEEE Transactions on Visualization and Computer Graphics*, 21(4):539–544, 2015.

[10] B. Cain. A review of the mental workload literature. Technical report, Defence Research And Development Toronto (Canada), 2007.

[11] K. Carpentier, D. Lourdeaux, and I. M. Thouvenin. Dynamic selection of learning situations in virtual environment. In *ICAART (2)*, pp. 101–110, 2013.

[12] C. M. Carswell, D. Clarke, and W. B. Seales. Assessing mental workload during laparoscopic surgery. *Surgical innovation*, 12(1):80–90, 2005.

[13] M. Cavazza, J.-L. Lugrin, D. Pizzi, and F. Charles. Madame bovary on the holodeck: immersive interactive storytelling. In *Proceedings of the 15th ACM international conference on Multimedia*, pp. 651–660. ACM, 2007.

[14] C.-J. Chao, S.-Y. Wu, Y.-J. Yau, W.-Y. Feng, and F.-Y. Tseng. Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 27(4):187–196, 2017.

[15] P. Chevaillier, T.-H. Trinh, M. Barange, P. De Loor, F. Devillers, J. Soler, and R. Querrec. Semantic modeling of virtual environments using mascaret. In *2012 5th Workshop on Software Engineering and Architectures for Realtime Interactive Systems (SEARIS)*, pp. 1–8. IEEE, 2012.

[16] G. Claude, V. Gouranton, R. B. Berthelot, and B. Arnaldi. Short paper:# SEVEN, a sensor effector based scenarios model for driving collaborative virtual environment. 2014.

[17] J. Collins, H. Regenbrecht, T. Langlotz, Y. S. Can, C. Ersoy, and R. Butson. Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 351–362. IEEE, 2019.

[18] J. Cremer, J. Kearney, and Y. Papelis. HCSM: a framework for behavior and scenario control in virtual environments. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 5(3):242–267, 1995.

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

[20] D. De Waard. *The measurement of drivers' mental workload.* Groningen University, Traffic Research Center Netherlands, 1996.

[21] A. Dey, A. Chatourn, and M. Billinghurst. Exploration of an EEG-based cognitively adaptive training system in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 220–226. IEEE, 2019.

[22] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with computers*, 21(1-2):133–145, 2009.

[23] L. Freina and M. Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The International Scientific Conference eLearning and Software for Education*, vol. 1, p. 133. " Carol I" National Defence University, 2015.

[24] S. Gerbaud, N. Mollet, and B. Arnaldi. Virtual environments for training: from individual learning to collaboration with humanoids. In *International Conference on Technologies for E-Learning and Digital Entertainment*, pp. 116–127. Springer, 2007.

[25] D. Gopher and E. Donchin. Workload: An examination of the concept. 1986.

[26] F. Grimm, G. Naros, and A. Gharabaghi. Closed-loop task difficulty adaptation during virtual reality reach-to-grasp training assisted with an exoskeleton for stroke rehabilitation. *Frontiers in neuroscience*, 10:518, 2016.

[27] K. S. Gurusamy, R. Aggarwal, L. Palanivelu, and B. R. Davidson. Virtual reality training for surgical trainees in laparoscopic surgery. *Cochrane database of systematic reviews*, (1), 2009.

[28] S. Haga, H. Shinoda, and M. Kokubun. Effects of task difficulty and time-on-task on mental workload. *Japanese Psychological Research*, 44(3):134–143, 2002.

[29] T. C. Hankins and G. F. Wilson. A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, space, and environmental medicine*, 69(4):360–367, 1998.

[30] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[31] K. F. Hew and W. S. Cheung. Use of three-dimensional (3-D) immersive virtual worlds in K-12 and higher education settings: A review of the research. *British journal of educational technology*, 41(1):33–55, 2010.

[32] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[33] R. Kizony, N. Katz, and P. L. Weiss. Adapting an immersive virtual reality system for rehabilitation. *The Journal of Visualization and Computer Animation*, 14(5):261–268, 2003.

[34] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, pp. 279–328, 1991.

[35] C. R. Larsen, J. L. Soerensen, T. P. Grantcharov, T. Dalsgaard, L. Schouenborg, C. Ottosen, T. V. Schroeder, and B. S. Ottesen. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. *Bmj*, 338:b1802, 2009.

[36] G. T. Leung, G. Yucel, and V. G. Duffy. The effects of virtual industrial training on mental workload during task performance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 20(6):567–578, 2010.

[37] T. Luong, N. Martin, F. Argelaguet, and A. Lécuyer. Studying the mental effort in virtual versus real environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 809–816. IEEE, 2019.

[38] Z. Merchant, E. T. Goetz, L. Cifuentes, W. Keeney-Kennicutt, and T. J. Davis. Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education*, 70:29–40, 2014.

[39] N. Meshkati and P. Hancock. *Human mental workload*, vol. 52. Elsevier, 2011.

[40] S. Miller. Workload measures. *National Advanced Driving Simulator. Iowa City, United States*, 2001.

[41] N. Moray. *Mental workload: Its theory and measurement*, vol. 8. Springer Science & Business Media, 2013.

[42] M. Narayan, L. Waugh, X. Zhang, P. Bafna, and D. Bowman. Quantifying the benefits of immersion for collaboration in virtual environments. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 78–81. ACM, 2005.

[43] R. O'Donnell and F. Eggemeier. Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. KR Boff, L. Kaufman and JP Thomas, 1986.

[44] T. D. Parsons and J. L. Reinebold. Adaptive virtual environments for neuropsychological assessment in serious games. *IEEE Transactions on Consumer Electronics*, 58(2), 2012.
[45] R. W. Picard. *Affective computing*. MIT press, 2000.
[46] J. Psotka. Immersive training systems: Virtual reality and education and training. *Instructional science*, 23(5-6):405–431, 1995.
[47] F. Putze, C. Herff, C. Tremmel, T. Schultz, and D. J. Krusienski. Decoding mental workload in virtual environments: a fNIRs study using an immersive n-back task. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3103–3106. IEEE, 2019.
[48] G. B. Reid and T. E. Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology*, vol. 52, pp. 185–218. Elsevier, 1988.
[49] D. Reinhardt, S. Haesler, J. Hurtienne, and C. Wienrich. Entropy of controller movements reflects mental workload in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 802–808. IEEE, 2019.
[50] A. H. Roscoe. Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological psychology*, 34(2-3):259–287, 1992.
[51] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr. The multi-attribute task battery II (MATB-II) software for human performance and workload research: A user's guide. 2011.
[52] N. E. Seymour, A. G. Gallagher, S. A. Roman, M. K. O'brien, V. K. Bansal, D. K. Andersen, and R. M. Satava. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery*, 236(4):458, 2002.
[53] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 3(2):130–144, 1994.
[54] M. B. Sterman, C. A. Mann, D. A. Kaiser, and B. Y. Suyenobu. Multiband topographic EEG analysis of a simulated visuomotor aviation task. *International journal of psychophysiology*, 16(1):49–56, 1994.
[55] N. Szilas. IDtension: a narrative engine for interactive drama. In *Proceedings of the technologies for interactive digital storytelling and entertainment (TIDSE) conference*, vol. 3, pp. 1–11, 2003.
[56] A. J. Tattersall and P. S. Foord. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5):740–748, 1996.
[57] D. Traum, J. Rickel, J. Gratch, and S. Marsella. Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 441–448. ACM, 2003.
[58] C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski. Estimating cognitive workload in an interactive virtual reality environment using EEG. *Frontiers in Human Neuroscience*, 13, 2019.
[59] P. S. Tsang and V. L. Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.
[60] P. S. Tsang and M. A. Vidulich. Mental workload and situation awareness. 2006.
[61] W. B. Verwey and H. A. Veltman. Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of experimental psychology: Applied*, 2(3):270, 1996.
[62] D. Waller, E. Hunt, and D. Knapp. The transfer of spatial knowledge in virtual environment training. *Presence*, 7(2):129–143, 1998.
[63] C. D. Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.
[64] W. W. Wierwille and F. T. Eggemeier. Recommendations for mental workload measurement in a test and evaluation environment. *Human factors*, 35(2):263–281, 1993.
[65] D. Wu, C. G. Courtney, B. J. Lance, S. S. Narayanan, M. E. Dawson, K. S. Oie, and T. D. Parsons. Optimal arousal identification and classification for affective computing using physiological signals: Virtual reality stroop task. *IEEE Transactions on Affective Computing*, 1(2):109–118, 2010.
[66] B. Xie and G. Salvendy. Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & stress*, 14(1):74–99, 2000.
[67] F. R. H. Zijlstra. *Efficiency in work behaviour: A design approach for modern tools*. PhD thesis, 1993.