The effect of Task complexity and Mental workload on Task performance in an immersive Virtual reality training

Lilibeth Juárez Rojas

Faculty of Behavioral Management and Social Sciences, University of Twente

Master Thesis

Supervisors:

Iliona Friso van den Bos

Bas Kolloffel

16/09/2021

Author Note

Lilibeth Juárez Rojas, Faculty of Behavioral Management and Social Sciences, University of Twente.

Correspondence concerning this article should be addressed to Faculty of Educational Science and Technology, University of Twente, Drienerlolaan 5, 7522 NB Enschede, the Netherlands.

Contact: lilibeth.juarez@gmail.com

Table of Contents

Abstract

There is an increasing interest to develop engaging, practical, and cost-effective training solutions for workers in the construction industry. Virtual Reality (VR) has been demonstrated to be a suitable solution to cope with those demands, but important factors must be considered before implementation. Task complexity (TC) and mental workload (MWL) are well-known predictors of task performance, and their effects should be evaluated to maximize VR effectiveness. This study investigates what is effect of different levels and order of TC and MWL on task performance to train future operators in the asphalting industry. The following main research questions are guiding the present study: *Are there significant differences in rates of performance when participants are assigned to perform in different levels and orders of TC?,* and *does MWL have a moderation effect on task performance of participants when they are assigned to perform under different scenarios of TC?* Research questions were answered by using the levels and order of TC as independent variable, performance results as dependent variable, and MWL measures as covariate. This investigation uses an experimental design based on quantitative data collection. Ten students from the University of Twente took part of the study. Task performance was measured using the performance scores from the VR system. TC had two levels: low and high. Heart rate measures were used to assess objective MWL and Self-perception surveys for subjective MWL. Data was analyzed with SPSS, using an ANOVA and ANCOVA Repeated Measures analysis. No significant differences were found among training groups. Starting with the low TC or with the high TC did not make a difference in terms of performance. Second, there was no evidence that MWL has an influence on task performance. The effect of objective and subjective MWL did not have a significant impact on performance results. This study contributes to further understanding of how the order and levels of TC and MWL might influence on task performance within an immersive VR environment.

Keywords: Task complexity, mental workload, virtual reality

The effect of task complexity and mental workload on task performance in an immersive virtual reality training

In recent years, there has been an increasing interest to develop engaging, practical, and cost-effective training solutions for employees in the construction industry. According to the European Construction Sector Observatory (2018), the Dutch construction sector risks a shortage of skilled workers, low numbers of students in subject-related areas and an increasing proportion of older workers nearing retirement age. As a result, important initiatives such as the Digital Education Action Plan (2021-2027) are focusing on strengthening the quality of instructional methods of vocational education training (VET) institutions, making them more practical and hands-on.

However, current and conventional construction training still includes traditional methods such as lecture presentations, video training, job shadowing or on-site training. This might result in unengaging or expensive training methods (Schwarze et al., 2019). Besides, construction is one of the most hazardous industries in which workers often have to make decisions under pressure or unexpected changes in their work routine. The nature of construction sites makes on-site training difficult and prevents training through the experience of failure. For instance, workers are not allowed to make mistakes in the real setting because this could have adverse consequences such as getting hurt or damaging the machinery. Therefore, it is recommended to use more experiential and natural settings to practice those changing and stressful conditions (Sacks, Perlman, & Barak 2013). In this sense, Virtual Reality (VR) is one of the most practical, safe, and cost-effective alternatives for learners to practice hazardous tasks that are difficult or impossible to emulate in the real-life (Gutierrez et al., 2017; Slater & Sanchez, 2016).

VR is a computer-based environment, wherein someone can move and interact with the elements in real-time by using a set of controls. The closeness to the real-life experiences is one of the main technical attributes of VR systems which is considered a key aspect to engage learners to improve the learning outcomes (Slater, 2003). Numerous studies have demonstrated its effectiveness to improve the learning outcomes in the construction industry (Goulding et al., 2012; Joshi et al., 2020; Rezazadeh et al., 2011; Sacks, Perlman & Barak, 2013; Vahdatikhaki et al., 2019). Besides, research has demonstrated that

VR interactive training environments provide means to get leaners to experience the training goals (Magerko, 2002), help support training transfer and accelerate training (Jarvis & Freitas, 2009). Indeed, construction training is the second largest application of educational VR after healthcare and has been rapidly recognized in construction engineering education and training programs since VR is believed to be an effective tool to enhance training programs (Wang et al., 2018).

In the context of dynamic systems such as VR, the conditions of training practice are usually evaluated to achieve better performance results without exceeding the learner's mental resources. Settings are traditionally designed to start with the novice and then with the expert level or with the simple and then with the most difficult task. It is usually believed that executing in the order from novice to expert allows individuals to improve task performance. However, varying the conditions of practice or providing unpredictable events might create difficult conditions that could enhance the performance outcomes (Bjork, 1994). Research has demonstrated that the introduction of a variety of motor, verbal and problem-solving tasks in the training environment improves the long-term performance specially to transfer training in novel and related task environments (Bjork, 1994; Del Rey et al., 1982; Goettl, 1994; Shea & Morgan, 1979; Hall, Dominguez & Cavazos, 1994; Young, Cohen & Husak, 1993;). This means that the introduction of difficulties during the execution of the training practice causes difficulties for leaner but improves performance in long-term. The latter might be tested through the order manipulation, unexpected changes or randomized ways to figure out how it impacts the performance outcomes (Bjork, 1994).

However, the introduction of such difficulties might cause for learners to demand more mental resources to perform the task. The variation of task complexity might require learners to use more skills, knowledge, cognitive abilities, memory capacities and task effort, demanding a higher human processing information, resulting in increased levels of mental workload (Jacko & Waard, 1996). An increased mental workload might exceed learners' mental resources, avoiding them to capture critical information for their safety. Thus, the impact of mental workload derived by the variations in task complexity during training is an essential factor that should be measured to evaluate its effect on users and to set VR as an effective training tool. Mental workload is widely recognized as an important factor to predict task performance in complex systems and training procedures (Carswell, et al., 2005; Dahlstrom et al., 2009; Gopher et al., 1986). To set VR as an effective training tool, there must be a balance between the imposed task demands within the VR environment and the learner's mental resources to achieve the better performance results.

Therefore, the present study aims to investigate the effect of different levels and order of task complexity on task performance as well as to test the moderator effect of MWL on task performance. This research is expected to contribute to the growing research that involves the use of physiological measurements to test the effect of MWL on task performance and where the order of task complexity is variated. The contribution of this work is novel, because to the best of the author's knowledge, this study would be one of the first of its kind to test the combined influence of task complexity and the effect of mental workload on task performance using immersive VR as a training tool in the asphalting construction industry in the Netherlands.

Virtual Reality training for the construction industry

From a technological perspective, Virtual Reality (VR) is defined as a computer-generated virtual environment that may be moved and manipulated by a user in real-time (Warwick et al., 1993). VR is a 3D simulation of the real world in which someone is represented through an avatar allowed to interact and manipulate the elements of the artificial environment. In terms of human experience, VR refers to "a simulated environment in which a perceiver experiences telepresence" (Steuer, 1992 pp. 76-77). Telepresence is the extent to which a person "feels present" in the artificial environment. For someone to "feels present" in the virtual world it depends on how much he or she feels immersed within it, and to feel immersed depends to what extend the individual is being absorbed or engaged with the simulated environment. In this respect, the term immersion is referred as the mental state of being completely engaged or absorbed with something, in which other demands are ignored (Agarwal & Karahanna, 2000; Dede, Jacobson, & Richards, 2017).

According to Slater (2003), the level of immersion depends on the physical attributes of the VR system. In the literature there are two types of VR generally accepted: non-immersive and immersive systems**.** Non-immersive VR refers to the systems that do not require highest levels of graphics performance, for example, simulations in desktop computers or projection screens (Zahabi & Abdul, 2020) and immersive VR refers to the technology in which the users are fully engaged in the artificial environment through using special hardware such as head-mounted displays (HMDs) and sensor gloves (Wang et al., 2018).

In terms of human experience, immersive VR systems have important advantages over desktop-based VR due to the technological attributes that they offer. For instance, VR based on HMDs such as Oculus Rift or HTC Vive provides immersive experiences created by images, sounds, or other virtual scenarios so that the user can feel the virtual world is authentic and genuine (Wang et al., 2018). These perceptions imply the use of high-flow mental states with technology that is considered an important and beneficial aspect to improve learning outcomes (Mills & Noyes, 1999). Having high-flow mental states with technology is closely related to the construct of cognitive absorption. Cognitive absorption is based on the concept of flow (Agarwal & Karahanna, 2000), that is the mental state of absorption, a feeling of engagement, a sense of being in control, a loss of self-consciousness, and a shift in perception of time (Csikszentmihalyi, 1996). As a result, these deeper mental states that immersive VR systems induce on learners may help them to have more engaging learning experiences enhancing their individual performance during training.

Another important characteristic of the VR-HMDs or immersive VR is that it allows users to interact, create, or manipulate objects in real-time by using the controls provided by the system. These interactions generate real-time actions interpreted and coordinated procedurally by the user augmenting its learning through experience (Psotka, 1995). That means immersive VR creates a compelling interaction between mind and body that allows learners to establish a sensory integration to execute and learn processes in real-time (Psotka, 1995). In fact, the sensory motor training helps the learner to apply the knowledge from the VR to the real task (Rose et al., 2000). In a review conducted by Martin and colleagues (2021), the authors suggest that using multimodal VR that contains visual, tactile, and auditory stimulus provides a higher user engagement leading to a better experience and learning transfer. Immersive VR contain all these capabilities to enhance the learning outcomes. In fact, VR environment supports learners to build better mental representations of the real environment allowing them to better acquire the training skills (Adhikarla, et al., 2014). Therefore, it can be assumed that the use of controls to execute operations in real time, the visuals, and the use of sounds could promote the learning outcomes in operational matters.

Moreover, the characteristics of novelty, realism, fantasy and interactivity that immersive VR offers over other games or simulators creates motivation and engagement among learners (Malone & Lepper, 1987) and a motivated and engaged learner is more likely to meet the learning expectations. Physical attributes of VR may facilitate intrinsic motivation for learners by allowing them to have experiences difficult or impossible to emulate in the real world, giving them, for example the experience to manipulate objects within the VR environment (Dalgarno & Lee, 2010). The manipulation of objects related to the work environment and the realism that immersive VR offers over other technologies might be advantageous to support learners to maximize their motivation and engagement during training and thus enhance performance results in the long-term.

There is a large number of studies that have demonstrated the effectiveness of VR over traditional training methods in the construction sector (Cheng & Teizer, 2013; Goulding et al., 2012; Juang, Hung, & Kang 2013; Joshi et al., 2020; Li, Chan & Skitmore 2012; Rezazadeh et al., 2011; Sacks, Perlman & Barak 2013; Vahdatikhaki et al., 2019). For instance, Sacks, Perlman and Barak (2013) evaluated the effects of a 3D immersive VR system to assess training effectiveness in construction sites. Half of the participants received a traditional classroom training method, and the other half were trained using a 3D immersive VR power-wall system. The results showed that the VR training method was the more effective to create engaging learning experiences than traditional methods especially in matters of safety and on-site concrete works. In the same line and more recently, similar results were obtained by Joshi and colleagues (2020) who demonstrated that using an immersive VR training approach allows to engage more learners, providing them with a better understanding of safety protocols in real-life experiences in the precast concrete industry. Same results were obtained by Goulding et al. (2012), Li, Chan & Skitmore (2012) and more recently in paving operations, Vahdatikhaki et al. (2019).

The mentioned literature suggests that VR has demonstrated to cause better learning outcomes over traditional methods due to its attributes, which enhance the problem-solving, spatial, and motor skills. Thus, it can be assumed that the use of an immersive VR system could provide more engage and motivational learning experiences for learners allowing them to achieve better performance outcomes during training practices.

Decision-making training and task complexity

Construction workers are frequently exposed to various risks in which unpredictable events might happens during their job routines. High risks and changing environments require employees to make appropriate decisions to avoid being injured or cause damage to critical machinery. According to Horswill et al., (2008), individuals with better cognitive and psychomotor capacities create better mental representations of hazards and as a result they are more capable to cope with them. Cognitive and psychomotor processes include decision-making capacities, attention, time to respond, constricts sensitivity and visual pursuit, which are main important aspects of hazard perception skills (Horswill et al., 2008; Sumer, 2011). In complex and dynamic sociotechnical systems, abilities of decision-making are critical for safety and effectiveness (Jenkins et al., 2011). As a result, decision-making abilities must be taught to prepare trainees to cope with unpredictable events allowing to improve their performance in the real setting.

One of the learning theories associated to decision-making activity is the instance-based learning theory (IBLT). The IBLT proposes that in a dynamic decision-making context, individuals learn by accumulation, recognition, and refinement of instances (Gonzalez, Lerch & Lebiere, 2003). In other words, learners use their accumulated knowledge to make decisions taking advantage of their prior knowledge. These learning experiences may prepare learners to execute better under unexpected events in their working routines. The IBLT represents the decisions based on experience ranging from the least to the most dynamic tasks (Gonzales & Dutt, 2011). The least dynamic tasks involve sequential decisions in which the environment and the individual's information changes over time as a result of previous decisions and the most dynamic tasks are characterized to change more spontaneously as a result of previous decisions, attempting to maximize gains over long term (Edwards, 1962). In the context of immersive VR, for example, the learner first will interpret the artificial environment, then they will identify the goal and finally they will choose the best option among others based on their previous experiences to achieve the task. The use of an approach of learning-based experience might be advantageous in a context of immersive VR where the experiences might be used to enhance learning and performance outcomes. Preliminary work for developing specifications for VR environments support that individuals' decision-making abilities are closely related to their experience in a particular domain and their rule-based heuristics (Jenkins et al., 2011). Hence, it might be expected that the introduction of more dynamic tasks could enhance performance during training.

To evaluate this aspect, many types of manipulations would take part of training programs to enhance the training decision-making abilities and consequently to improve the performance outcomes. According to the desirable difficulty framework, the introduction of difficulties and challenges in a training routine might enhance performance results (Bjork, 1994). Varying the conditions of practice (i.e. practice trials in a random fashion), providing contextual interference (i.e. unpredictable events), distribution of practice and reducing feedback might create conditions to enhance the long-term performance (Bjork, 1994). For instance, contextual interferences and the random practice are difficulties that might be included in training environments to maximize the performance outcomes. According to Battig (1979) the introduction of contextual interferences and varieties produces more elaborate and distinctive learning which might result in better retention of information, and consequently in better transfer at the time of retrieval. In the same context, scheduling the practice trials in a random way has been shown to impair performance during training enhancing long term performance (Shea & Morgan, 1979; Hall, Dominguez & Cavazos, 1994). In the construction industry, the introduction of these types of manipulations might be advantageous for effective training since unexpected events could happen during work routines (i.e. technical problems, degree of compaction required, environmental conditions, etc). These types and levels of difficulty might be represented by the construct of task complexity.

Task complexity is defined by the elements that compose a particular task, the relationship among those elements, and the learner expected behaviour, the quantity, interaction, and variation of those elements establish the complexity to perform a desired task (Campbell, 1988; Wood, 1986). The complexity of a task is determined by the incorporation of individual attributes (e.g. high or low) and by the total number of attributes of the task (Campbell, 1988). In practice, tasks are designed as low or simple and high or complex, but low or simple tasks are always less demanding than high or complex tasks. The fixed features of tasks demand different levels of cognitive resources to achieve the desired task (Robinson, 2014). This means that individuals recall their cognitive resources in function of the complexity of the task.

However, learners might perceive task complexity in different ways being determined by the characteristics of the task, and their self-perceptions. Self-perceptions and cognitive factors are contributors of task complexity to predict task performance (Mangos & Steele-Johnson 2001). This is referred as task difficulty and varies from person to person. According to Peng & Zhizhong (2011), the structure of the task imposes certain resource demands on the performer which are mediated by their perception of complexity and by the total amount of resources brought by them to accomplish the required task. That means that the difficulty perception of the task and the characteristics of the task itself determine and predict task performance.

In the literature, only a few studies have investigated such effect obtaining positive results in military and industrial settings respectively (Asare & McDaniel, 1996; Marshall & Byrd, 1998; Mascha, 2001; Pepinsky et al., 1960; Rao et al., 2019). For instance, in a recent study, Rao and colleagues (2019) tested the effects of a 3D VR training varying the order of task difficulty in the military domain. Half of the participants executed the task in a novice scenario first and the other half executed the expert scenario first. The results showed that participants who did the expert scenario first performed better than the participants that executed the novice scenario first. The latter is a nice example of how varying the conditions of practice can result in better performance. Starting with the expert level created better mental representations for learners and as a result they obtained better performance results.

Thus, according to the IBLT and the desirable difficulty framework, the learner will obtain superior experiences when the complexity of the task is high. Therefore, it is expected that the accumulated experiences of the high task complexity allow learners to obtain better mental representations of the imposed tasks which eventually might result in enhanced decision-making abilities and consequently better performance. Variating the levels of complexity (low and high) through the conditions of practice (randomized way) and unpredictable events (i.e. weather conditions) might improve the performance of learners with decision-maker roles.

Mental workload and task performance

In the literature the term mental workload refers to the mental resources used to execute a task. Mental Workload (MWL) is the total amount of mental-processing resources used to accomplish a task, resulting in a decrement of the human internal sources (Waard, 1996; Wickens & Hollands, 2000; Xie & Salvendy, 2000). In other words, the mental resources are used to accomplish a task that might eventually result in stress. In construction industry, measuring the impact of mental workload is crucial, since in roadway work zones, a driver could experience high task demands and if it exceeds their mental resources, operators might fail to capture critical information for their own safety (Shakouri et al., 2018). In fact, it has been supported that optimizing operators' mental workload could reduce errors, improve safety, increase productivity and enhance operators' satisfaction (Cain, 2007; Moray, 2013; Roscoe, 1992; Tsang & Vidulich, 2006).

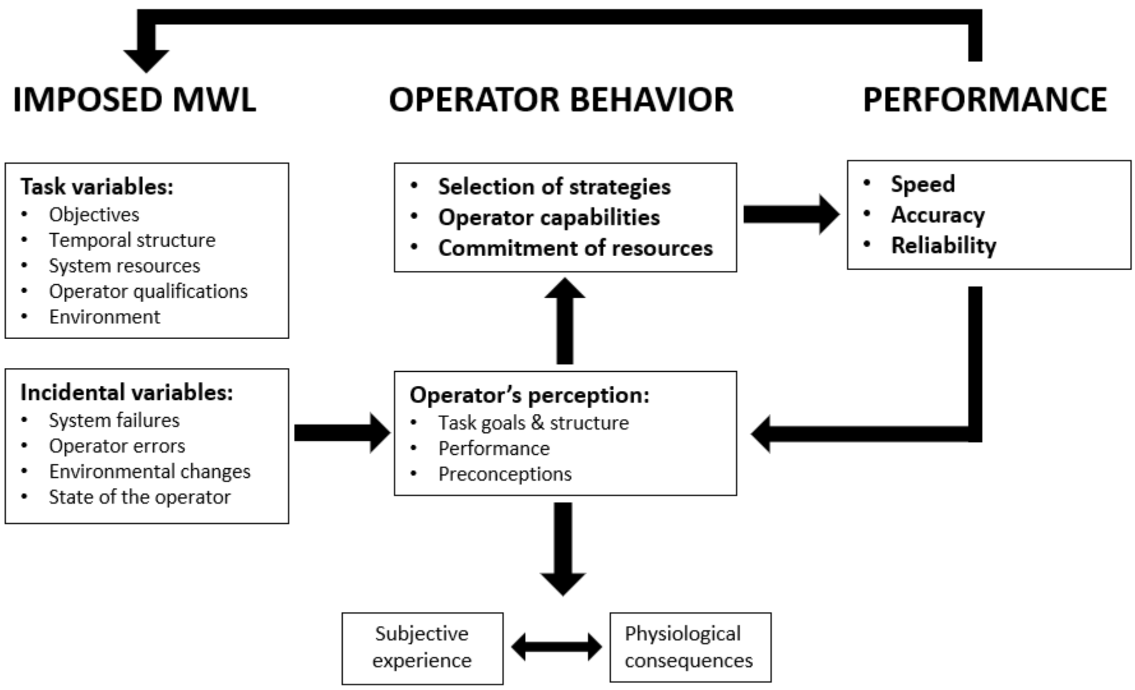
Likewise, in complex systems such as immersive VR, mental workload is an important aspect used to measure the extent that a cognitive activity exceeds the user's mental resources to execute a particular task. A VR system can be optimized considering mental workload levels at early-stage of design to guide designers to make appropriate adjustments (Xie & Salvendy 2000). Mental workload has been widely recognized as an important factor to predict task performance in complex systems and training procedures (Carswell, et al., 2005; Dahlstrom & Nahlinder, 2009; Gopher & Donchin, 1986). In this sense, mental workload has been studied from two perspectives; the first one sustains that mental workload depends on task demands to which the individual adapts and the second one considers that mental workload is a consequence of the relationship between task demands and the performer's skills in terms of the balance between demand and resource (Ferrer & Dalmau, 2004; Young & Stanton, 2005). The last perspective has received more support.

The latest approach suggests that mental workload emerges from the interaction between the structure of the task, the environmental conditions, the skills, behaviours, and perceptions of the operator (Hart & Staveland 1988; Young et al., 2015). Mental workload's meaning depends on learners' experience, expectations, and understanding of task demands (Hart & Staveland, 1988). This implies that mental workload is composed by the task demands and how the learner perceives the imposed demands and environmental conditions. The demand imposed by the task, operator's subjective mental workload and other influences may combine to create overall mental workload.

In this respect, Hart & Staveland (1988) proposed a conceptual framework for relating variables that influence mental workload on performance (Figure 1). In this model, performance is influenced by the imposed mental workload that consist of the task variables and incidental variables. The latter, have an influence on the operator's perception of task, that is reflected as physiological changes and subjective experiences. The operator's perception determines their behavior and consequently has a direct impact on performance. Thus, it can be assumed that structure of the task, the environmental conditions wherein the task is executed, the capabilities of the users and their perceptions are the main factors that will determine mental workload and consequently task performance results.

Figure 1

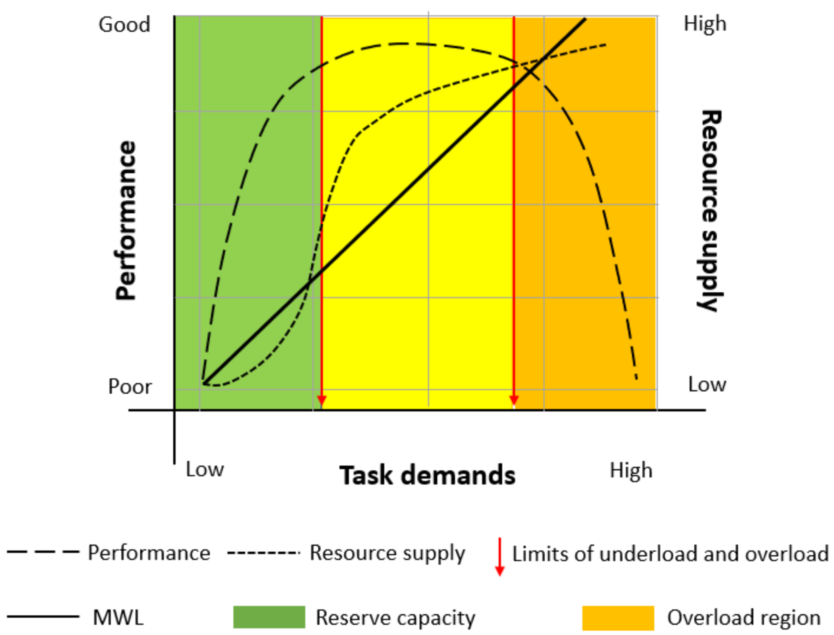
Conceptual framework for relating variables that influence MWL and performance



However, the complexity of the VR environment might affect the information processing resulting in increased or decreased mental workload. In the case of task complexity, Jacko and Ward (1996) posit that a high-task complexity demands more skills, knowledge, cognitive abilities, memory capacities and task effort, demanding higher human processing information resulting in increased levels of mental workload. Higher levels of mental workload cause overload that is when an operators face more stimuli than they can handle, and the excessive load could affect their selective attention, leading them to have lower levels of performance (Easterbrook, 1956). The same effect on performance can be obtained in the contrary case, when the imposed task causes little stimulation which produces an underload effect. The underload effect happens when the mental resources are assigned to another subject or are reduced due to its underuse (Young & Stanton, 2002). In this sense, Young et al. 2015 proposes a theoretical representation of such effect (Figure 2).

Figure 2

The supply-demand relationship associated with MWL and performance



In this model, the horizontal axis represents the task demands from low to high, and the vertical axis represents the performance from poor to good and the resource supply from high to low. The redline is the break point on the performance curve, which divides the three regions of the supply demand. The resource supply refers to the supply of available mental resources to accomplish a task. When task demands exceeds the resource supply, further demand increases, leading to decrement in performance.

In figure 2, the green area is the reserve capacity, that is the area in which mental workload increases and leads to improvements in performance as more resources are used to meet the increasing demand. In the central or yellow area, the workload gradually increases while performance is at its best, remaining relatively constant. Finally, in the overload region or orange area, the increase of mental workload results in lower performance. As a result, the reserve capacity and the overload regions have strong inference to predict and assess mental workload. The overload and underload effect are widely recognized and documented that can negatively affect performance (Xie & Salvendy, 2000).

Considering the desired difficulty framework which argues that the introduction of obstacles, variations and unexpected events produces difficulties for the leaners but enhances performance (Bjork, 1994), and the perspective of Jacko and Ward (1996), where a high task complexity increases the demand in cognitive abilities, memory capacity, task effort and mental resources resulting in an increased mental workload; It is expected that the introduction of obstacles and the variated order of task complexity lead to a higher mental workload and higher performance in consequence. In other words, better performance will be obtained under the high task complexity conditions within an acceptable range of mental workload. Such levels of mental workload are expected to be reflected via physiological responses and subjective experiences as the model of Hart and Staveland (1988) illustrates.

Mental workload measurements

The literature recognizes three main methods to measure mental workload: performance measures, subjective measures and physiological measures. Performance measures are based in final scores, speediness, time completion and number of errors to evaluate workload with respect to the changes in the task (Shakouri, et al., 2018). One of its main advantages is that the total amount of performance might indicate the total amount of workload in the user (Ogden et al., 1979; Wang, 2012). However, the insensitivity to the state and condition of the user is one of its disadvantages (Shakouri, et al., 2018). Therefore, this technique is usually supported by subjective and physiological measures.

To measure subjective mental workload, users execute the task and provide feedback on their workload perceptions. For instance, the NASA Task Load Index is a well-known and widely accepted tool for measure mental workload. This tool rates the perceived mental workload to assess the task, a system, or any aspect of performance (Hart & Staveland, 1988). The simplicity and inexpensiveness to collect users' perceptions are the main advantages of this tool. However, data collection could vary on time because is impossible to gather information in real-time and users'perception might be biased due to other factors than workload (Shakouri et al., 2018). Thus, subjective measures could be supported by physiological measures.

The response of the body to external stimulus is commonly used as indicators of mental workload (Waard, 1996). Some of the most frequent physiological signals used to measure mental workload are the cardiac, electrodermal, and brain activity. However, cardiac activity is the most common method used to measure mental workload. The physiological indicator of Heart Rate (HR) is a well-known indicator of stress and mental workload (Healey & Picard, 2004). An increased heart rate usually is reflected when subjects have to expand their mental efforts to execute a task in comparison to resting situations (Mulder, Waard & Brookhuis, 2005). Many mental-effort studies are characterized by an increase of heart rate, which is considered as a defense reaction using short-lasting tasks that require challenging mental demands in working memory (Mulder, Waard & Brookhuis, 2005). Moreover, previous research shows that higher stress usually leads to higher heart rate (Abbe et al., 2011).

One of the main advantages of using physiological signals to measure mental workload is that the results are representative of the actual task workload. However, in the literature mixed results have been reported about the theory behind the measurement of physiological signals that have been not fully developed or has certain epistemic uncertainty (Casner & Gore, 2010). Therefore, this study uses the three methods to effectively collect objective and subjective measures of mental workload.

Current research

There is large and increasing amount of research that has been studied the constructs of mental workload and task performance in the industry (Chao et al., 2017; Das, Maiti & Krishna 2020; Leung, Yucel & Duffy, 2010; Shi et al., 2020), in the driving settings (Fan, 2018; Heikoop et al., 2019; Michaels et al., 2017; Shakouri et al., 2018; Schiessl, 2008), and military field (Lackey et al., 2016; Luong et al., 2020; Mansikka et al., 2016; Perry et al., 2008; Sakib et al., 2020) within a context of VR training environments. Besides, there is an emerging approach to design and evaluate computer-systems based on learning algorithms to predict mental workload on task performance as Jebelli (2019) and Longo (2018) studies have been demonstrated.

However, only a small number of studies have considered the variation of order of task complexity and the effect of mental workload on task performance using an immersive VR training environment. Some of them using subjective tools to measure performance (Leung, Yucel & Duffy, 2010; Luong et al., 2020), physiological measures to assess mental workload (Shi, 2020), or both (Chao et al., 2017; Das, Maiti and Krishna, (2020); Sakib et al., 2020; Shakouri et al., 2018). For instance, the study conducted by Chao and colleagues (2017), evaluated the effects between a VR training and a traditional training method (technical manuals and multimedia films) on performance and mental workload, using a simple and complex task. Objective MWL was measured by using galvanic responses and cardiac activity, and a NASA-TLX questionnaire to measure subjective MWL. The results of the performance measures showed that the VR training method was considered the best to execute complex tasks compared to the traditional training approach. The physiological measures showed that there were significant differences between both training methods, wherein the VR training method produced a lower objective and subjective mental workload results. In contrast, Mansikka et al., (2016) found significant differences in physiological measures when task demands were varied in a simulated flight environment. The authors compared heart rate variability and performance under different proficiency tests. The results demonstrated that pilots showed different levels of heart rate variability across the different scenarios of complexity but maintained a high and mostly equal rates of performance. Shakouri et al., (2018) found a similar effect on performance, but the heart rate measures were not affected by the traffic densities in a driving scenario. The results showed that there was no relationship between subjective workload, physiological workload, and driving performance indicators. Sakib and colleagues (2020) also reported a similar effect using an immersive VR training for drone operators.

Overall, these studies measured the mental workload derived from the different scenarios imposed as a commonly factor to evaluate VR systems. The objective and subjective MWL measures are methods widely accepted. Most of the studies mentioned above used both objective and subjective measurements to create reliability in results (Chao et al., 2017; Das, Maiti and Krishna, (2020); Sakib et al., 2020; Shakouri et al., 2018). However, the results are mixed and varied due to the different environments and conditions of practice. In some cases, there were in favor of VR and better performance outcomes were obtained compared to traditional training methods (i.e. Chao et al., 2017), in other cases there were no differences in performance, but was difference in mental workload measurements (i.e. Mansikka et al., 2016), and conversely in other cases there was no relationship between mental workload and performance when the training was executed in the real practice (i.e Sakib et al., 2020).

Although the existing mentioned literature examines task complexity and mental workload in VR environments, the above-mentioned studies are only focused on the effects that task complexity produces on performance results, considering mental workload as an indicator of performance and mostly focused on technical aspects from the VR systems or to test computer-systems based on learning algorithms to predict mental workload. Hence, to the best knowledge of the author, there is a relatively small body of literature that is concerned with the effect that produces the levels and order of task complexity as well as to test the moderation effect that mental workload might have on task performance, without considering it as an indicator of performance within an immersive VR environment in the asphalting construction industry.

Therefore, the present study has two main purposes: 1) investigate whether the introduction of different levels of task complexity in a variated order has a positive effect to enhance performance outcomes, and 2) to determine the extent to which mental workload has a moderation effect on task performance. Thus, to examine the way in which task complexity and mental workload have an impact on task performance, this study is subdivided into 4 research goals: 1) whether the introduction of difficulties in the high task complexity causes significant differences among training groups, 2) whether the order of task complexity has a significant effect in terms of performance among training groups, 3) whether there are significant progress among practice trials between training groups, and 4) whether mental workload has a moderation effect on task performance results.

Hence, to test whether there are significant differences in terms of task performance, between the training groups that executed under different levels of task complexity, the first research question is: *Are there significant differences in rates of performance when participants are assigned to perform under different levels of task complexity?*

It is hypothesized that the introduction of difficulties contained in the high task complexity causes better performance results, in comparison with the low task complexity. Therefore, it is expected that there were significant differences in terms of performance among levels of task complexity.

Then, to research whether the order of task complexity has a significant influence on the performance results, the second research question is: *Are there significant differences in rates of performance when participants are assigned to perform under different order of task complexity?*

It is hypothesized that the order of task complexity has a significant effect on the performance results. It is expected that there were significant differences in performance outcomes between the training group that first executed the high task complexity condition and the training group that first encountered the low task complexity.

Next, to examine if there was a significant progress between the first and second trial of task complexity, the third research question is the following: *Is there a significant progress of participants' performance rate between the first and second trial of task complexity?*

It is expected there was a significant progress between the first and second trial for both training groups.

Lastly, to research whether mental workload has a moderation effect on task performance, the fourth research question is: *Does mental workload have a moderation effect on task performance of participants when they are assigned to perform under different scenarios of task complexity?*

It is expected that mental workload has a positive moderator effect on task performance. The high task complexity will produce increased levels of mental workload that would produce better performance results.

Method

Research design and participants

This project used an experimental research design aimed to investigate the effect of task complexity and mental workload on users' task performance in an immersive VR training. For this purpose, a random sample of 14 students with different nationalities, pertaining to Engineering and Educational Sciences faculties from the University of Twente (UT) were recruited. In total 10 participants (7 women and 3 men) with ages ranging from 20 to 41 (*M* = 32.50 years, *SD* = 5.33) completed the session. The other 4 participants presented symptoms of motion sickness, or their measurements were not saved correctly, so it was decided to not consider their participation. The sample size was 10 participants with no inclusion criteria with respect to the domain of expertise driving, use of VR or the compacting process. However, 58.33% reported having driving experience (7-10 years), 58.34% reported having used VR 1 to 3 times and 41.66% reported not having used VR technology.

A convenience sample was selected to approach participants. This method was chosen due to its nature of practicability, especially considering the current covid-19 pandemic. The benefit of this method is the permission of more accessibility, ease, and speed to collect data compared to other methods (Bornstein, Jager, & Putnick 2013). Thus, an attractive announcement shared via social media network was utilized to recruit candidates from the UT, and at the end of each session, the participants were rewarded with a gift thanking them for their participation.

Instruments

Task performance measurement

To measure objective task performance, the scores resulting from the VR application were used as output indicators. The program was coded to deliver performance scores for each participant at the end of each simulation. In this project, task performance represents compaction performance. Compaction performance was the area asphalted and the number of the roller machine passes in the simulated road made by each participant under the two scenarios of task complexity.

To calculate compaction performance in overall, the scores were divided into two types: compaction performance and compaction quality. Compaction performance was calculated taking as a base the total results bigger than 0 and the total amount of spaces that should be covered (240 cells in total), in other words, was the percentage of the area compacted by the user. The calculation of compaction quality was based on the total amount of squares with results bigger than 1, divided by the total amount of squares to compact. It was the percentage of area compacted more than once by the user.

Objective mental workload measurement

To measure objective mental workload, physiological measures were collected by using the Empatica E4 wristband. The Empatica E4 is a research device that allows gathering in real-time physiological metrics like temperature, electrodermal activity, and cardiac activity (Empatica, 2020). The Empatica was used to capture heart rate measures and determine participant’s emotional arousal while undergoing the scenarios of task complexity wearing the VR system. This study utilized the mean of the heart rate measures to calculate objective mental workload in the low and in the high task complexity.

Objective mental workload was calculated using the difference between the mean of heart rate measures in the first trial and the mean of the heart rate measures in the second trial. The wristband detects the intensity of the light refracted on the skin caused by the fluctuations in blood flow (Empatica 2020), and an increased heart rate usually is reflected when subjects must expand their mental efforts to execute a task in comparison to resting situations (Mulder, Waard & Brookhuis, 2005). A hear rate increase may indicate that the user is under stressful conditions (Salgado et al., 2018). Many mental-effort studies consider an increase of heart rate as a defense reaction during short-lasting tasks that require challenging mental demands in working memory (Mulder, Waard & Brookhuis, 2005). Thus, the mean of heart rate was used as an indicator of mental workload in both task complexity conditions.

Subjective mental workload measurement

A semi-structured questionnaire was used as an instrument to collect subjective self-perception of mental workload and task performance, as well as to collect demographics such as age, gender, experience to drive, and familiarization with VR technology (Appendix A). The design of this form was based on the NASA TLX questionnaire (Hart & Staveland, 1988) that is commonly used to measure subjective levels of mental workload and performance. The instructions included marking the scale that participants felt most accurately representing their mental workload and performance perceived in the low task complexity scenario and the high task complexity scenario.

To collect data regarding the mental workload self-perception, the next queries were used: 1) On a scale of 1 to 5 please rate your level of stress experienced during this VR training in sunny conditions and 2) 1) On a scale of 1 to 5 please rate your level of stress experienced during this VR training in rainy conditions. The options to answer used a Likert scale ranging from 1 to 5 with the anchors of Not stressful, Moderately stressful, Considerably stressful, Very stressful, and Extremely stressful. In the case of performance self-perception during the execution of the two scenarios the following items were used: 1) How do you rate your performance under sunny conditions?, and 2) How do you rate your performance under rainy conditions?. The answers used a 10-point Likert scale, ranging from 1 to 10 with the anchors from Bad to Perfect.

Procedure

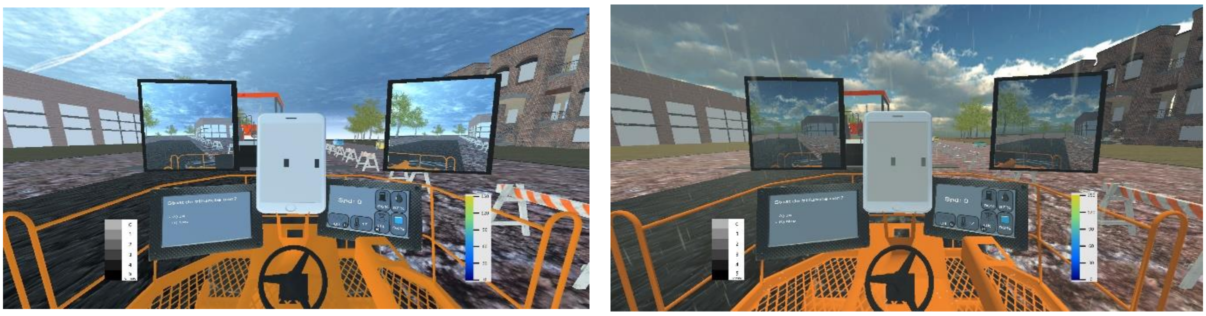
Prior to starting with the series of practices, ethical permission was requested from the Ethics Committee at the UT. Once the endorsements were approved (201138), the participants were recruited and scheduled to participate in the series of experiments. At the same time, the original PC-based application was adapted for Oculus Rift Suit VR and ran out with Unity 3D v2020. Then, the VR system was updated, ensuring its correct operation to run the application in a computer of the BSM Lab in the UT. Likewise, the E4 Empatica Manager was installed and tested to correctly capture the physiological measurements of the participants. Before each set of practices, all the equipment was sanitized, and its proper function was ensured.

In order to execute the practices, the participants were organized in a counterbalanced manner to explore the effect of order of task complexity on performance results and to get a clean comparison estimation among the training groups. Before coming to the practice, all participants were informed about the mandatory sanitary measures to prevent covid-19 in the lab and questioned about if they had symptoms, possible symptoms or if they had been in close contact with people infected with Covid-19. If all went well, their assistance was approved, and they could come to execute their practice. Then, each participant was received under a planned schedule previously agreed with them.

Before starting with the practice, a general explanation about the project was provided, wherein participants were allowed to ask questions about any aspect of the study. Then, the informed consent was given to the participants to read and sign it, and after that, the practice started. Task complexity consisted of a simulation of a roller operator's work routine under a sunny and rainy weather conditions. The sunny workday simulation represented the low-task complexity, and the rainy weather simulation represented the high-task complexity (Figure 3). The rainy condition contained the visual and audio effects that are expected to add difficulties for the learner and thus demand more mental resources to cope with compaction task.

Figure 3

Levels of task complexity



Low-task complexity (sunny weather conditions) High-task complexity (rainy weather conditions)

The training session consisted of an explanation of the main task to perform, how to operate the roller machine, and how to move between the options using the VR system controls using a power point presentation (Appendix B). Once assured that there were no questions and the participant felt ready to perform the required task, they were required to put on and adjust the Oculus Rift and the Empatica. Afterwards, devices were turned on, ensuring the synchronization between the E4 wristband and the management software. The participant was allowed to explore the default virtual reality scenario to become familiar with the virtual environment. Then, when the user felt ready and adapted in the immersive environment, the first simulation in the corresponding counterbalancing order was carried out. The participant did not know the order in which he or she would execute first.

Each simulation started with a simple questionnaire that asked the participants to review the driving indicators (i.e., water, gas, oil levels). Once answered, the virtual roller machine was enabled to turn on, and the participants could then begin the main compaction task. During this stage, participants were allowed to drive and compact as much as they could. This session lasted approximately 6 minutes, and then the simulation ended. Then there was a pause to save the performance measures and to ask participants about their comfort. If the participant did not feel comfortable the session ended. Otherwise, the second simulation was performed.

The second simulation was executed in the same way as the previous one: first answer the questionnaire, perform the main task, and collection of data scores. Once data records were properly saved, the participant was asked to get off the equipment and move to another desk. Last stage of this process consisted in answer the two items questionnaire, in which participants were required to self-assess their mental workload experienced during each of the VR sessions and performance during each scenario.

Data preparation

First, the performance scores provided by the VR system were converted to Excel files to calculate compaction performance. The application delivered the score results in a text file format in which the numbers represented the area and times of compaction in the simulated path. Those results were imported to an Excel file in form of matrix per each participant to calculate the total area and quality of compaction made in each scenario. The figure 4 illustrates an example of the performance scores obtained during first and second trial executing under the order sunny-rainy. The green area represents the area asphalted and the numbers inside are the times that the roller machine passed. Then, with an arithmetic formula in Excel, the percentage of compacted area and its quality was calculated.

Figure 4

Calculation of area compacted and quality compaction

				Sunny				
0	0	0	1	2	2	2	1	0
1	0	0	1	2	2	2	1	0
2	0	0	1	2	1	2	1	0
3	0	0	1	2	2	2	1	0
4	0	0	1	2	2	2	1	0
5	0	0	1	2	2	2	1	0
6	0	0	1	2	2	2	0	0
7	0	0	1	2	2	2	0	0
8	0	0	1	2	2	2	0	0
9	0	0	1	2	2	2	0	0
10	0	0	1	2	2	2	0	0
11	0	0	1	2	2	2	0	0
12	0	0	1	2	2	2	0	0
13	0	0	1	3	2	2	0	0
14	0	0	1	3	2	2	0	0
15	0	1	1	3	2	2	0	0
16	0	1	1	3	2	2	0	0
17	0	1	1	3	2	2	0	0
18	0	1	1	2	2	2	0	0
19	0	1	1	2	2	1	0	0
20	2	1	2	3	2	1	0	0
21	1	1	2	3	3	1	0	0
22	0	0	1	3	3	2	0	0
23	0	0	0	4	4	3	0	0
24	0	0	0	4	4	2	0	0
25	0	0	0	3	4	4	0	0
26	0	0	0	4	5	2	0	0
27	0	0	0	4	4	1	0	0
28	0	0	0	4	4	1	0	0
29	0	0	0	4	4	1	0	0
	% area covered				53.3333			
	% quality (more than 1)				35.8333			

				Rainy				
0	0	0	1	1	2	2	1	0
1	0	0	1	1	2	2	1	0
2	0	0	1	1	2	2	0	0
3	0	0	1	1	2	2	0	0
4	0	0	1	2	2	2	0	0
5	0	0	1	4	2	2	0	0
6	0	0	1	3	2	2	0	0
7	0	0	1	3	2	2	0	0
8	0	0	1	3	2	3	0	0
9	0	0	1	3	2	0	0	0
10	0	0	1	3	2	0	0	0
11	0	0	1	3	2	0	0	0
12	0	0	1	3	2	0	0	0
13	0	0	1	3	2	0	0	0
14	0	0	1	3	2	0	0	0
15	0	0	1	3	2	0	0	0
16	0	0	1	3	2	0	0	0
17	0	0	1	3	2	0	0	0
18	0	0	1	3	2	0	0	0
19	0	0	1	3	3	0	0	0
20	0	0	1	3	3	0	0	0
21	0	0	1	3	3	0	0	0
22	0	0	2	3	3	0	0	0
23	0	0	3	3	3	0	0	0
24	0	0	2	3	3	0	0	0
25	0	0	2	3	3	0	0	0
26	0	0	2	3	3	0	0	0
27	0	0	1	3	3	1	0	0
28	0	0	2	4	4	1	0	0
29	0	1	3	4	5	1	0	0
	% area covered				43.75			
	% quality (more than 1)				30			

To extract the physiological measures of mental workload, the heart rate measures from the Empatica Manager Software were downloaded to a computer. The pre-processing data analysis included a timestamp process to match the mean of heart rate with the time in which the scenarios were presented in the counterbalancing order (i.e. first sunny, then rainy and vice versa). The results of the heart rate mean per scenario were imported and organized per participant in an Excel file.

Once, participants' information was organized in an Excel file including their performance measures and mean of heart rate experienced during each scenario, data was pre-treated to eliminate outliers. Then, data from Excel was imported to an SPSS dataset to execute the further analyses.

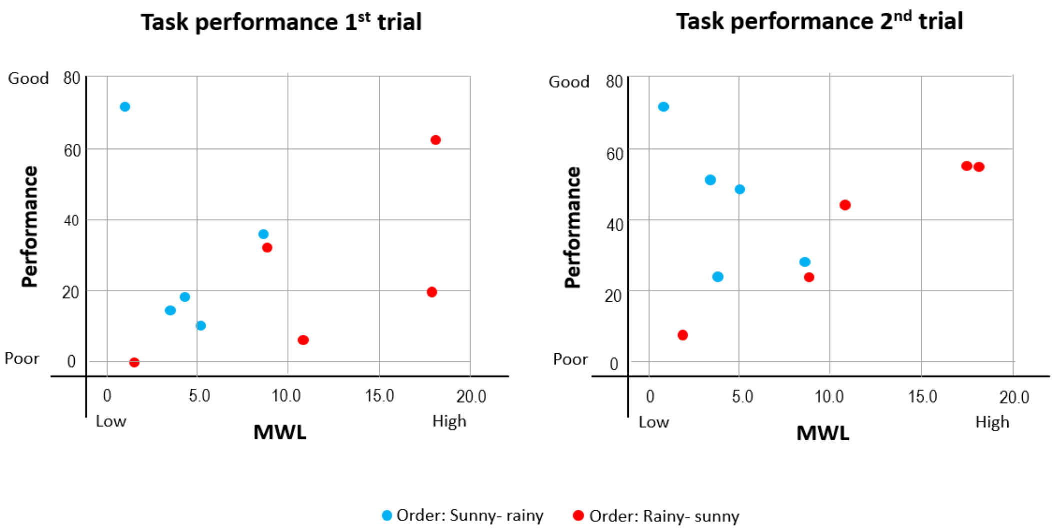
First, Normality was verified, Shapiro-Wilk test showed a normal distribution among the variables of compaction performance-first trial (order 0 [p = .151]; order 1 [p =.576]), compaction performance-second trial (order 0 [p = .739]; order 1 [p =.376]), and mental workload (order 0 [p = .151]; order 1 [p =.576]).

Next, the statistical assumptions of ANCOVA Repeated Measures were verified. Sphericity assumption was met due to the only two levels of the variable order (0 = start with sunny and 1 = start with rainy). A repeated measures variable with only two levels met perfectly the assumption of sphericity because the estimated computed by SPSS are 1, which is the perfect sphericity (Field, 2016). Likewise, the assumption of homogeneity of variance was met. The Levene's test showed an equal variance between the dependent variables (performance-first trial (p = .751; performance-second trial (p = .119).

In contrast, the assumption of the linear relationship between the covariate and the dependent variables was partially met. There was a partial linear relationship between the mental workload and the performance for the first and the second trial (Figure 5). The significance of the relationship between the mental workload (the covariate) and the order (the independent variable) was p = .012, a value less than 0.5, which means that, the assumption of homogeneity of regression slopes was violated. However, the researcher decided to execute the further statistical analysis to examine if there were significant differences among training groups and if mental workload had an influence on task performance results.

Figure 5

The partial linear relationship between MWL and the performance measures between trials.



Analytical strategy

To accomplish the purpose of this study that is investigating the effects of task complexity and mental workload on task performance among training groups, an Analysis of Covariance (ANCOVA) Repeated Measures was used. To address the first, second and third research questions, that is testing the difference in performance rates between the training groups where the order of task complexity was varied and to examine the rate of progress among groups, an ANOVA Repeated Measures was used.

Then, to answer the fourth research question which has as a purpose to examine whether mental workload has a moderation effect on the task performance, an analysis of covariance (ANCOVA) was utilized, using the difference between the mean of heart rate experienced by the participants during the first trial and the mean of Heart rate experienced during the second trial as Covariate. The heart rate difference was used as an indicator of the extent to which mental workload moderates the relationship between order of task complexity and task performance results.

As it was mentioned before, the analyses were supported by the counterbalancing technique. This technique allows dealing with the effects of the order when a repeated design is used (Field, 2016). Therefore, the sample was divided in half, with one half completing the task of compacting in one order (first sunny, then rainy) and the other half completing it in the reverse order (first rainy, then sunny).

This design uses as independent variables: order and task complexity. Task complexity has two levels of difficulty: low and high. The sunny condition was used to represent the low task complexity that is referred as the novice conditions, and the rainy condition was used to represent the high task complexity referred as the expert condition. Dependent variables consist of the task performance results given by the VR system and the results of the two self-perception items. Task performance scores use an interval scale and the results of the two self-perception items use an ordinal scale.

Results

Descriptive statistics

The mean and standard deviation of performance measures of the first and second trial in the different orders of task complexity are shown in Table 1.

Table 1

Descriptive Statistics of Performance measures in the different orders of task complexity

		M	*SD*	*N*
Performance	0 (Sunny-Rainy)	30.24	25.08	5
1st Trial	1 (Rainy-Sunny)	23.99	24.75	5
	Total	27.12	23.72	10
Performance	0 (Sunny-Rainy)	61.16	23.75	5
2nd Trial	1 (Rainy-Sunny)	45.41	30.12	5
	Total	53.28	26.88	10

The effect of task complexity on task performance

To assess the effects of task complexity and order variation on task performance, a Repeated Measures ANOVA was used. The analysis used task performance as dependent variable. The dependent variables of performance first trial and performance second trial as a within-subjects factor, and the order as a between-subjects factor (order 0= sunny-rainy, and order 1=rainy-sunny).

To respond to the first research question that is: *Are there significant differences in rates of performance when participants are assigned to perform under different levels of task complexity?*, the results of the test between-subjects factor showed that there was non-significant difference in terms of task performance, between the high task complexity and the low task complexity, F = .599, p = .461, partial η^2 = .070. This means that there were not significant differences of task performance between the sunny and the rainy conditions.

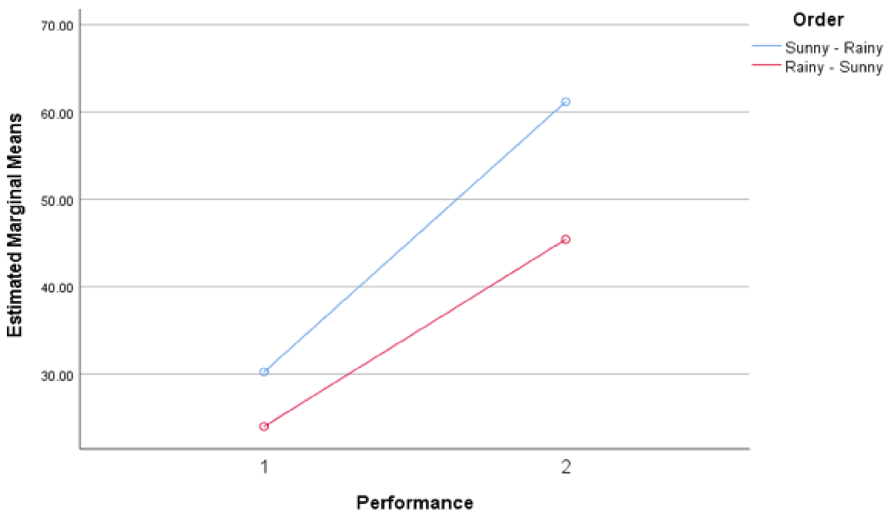
Likewise in order to respond to the second research question that is: *are there significant differences in rates of performance when participants are assigned to perform under different order of task complexity?*, the results of the test of within-subjects effects revealed that the interaction between performance and order was not significant. There was not a significant interaction between performance and order F = .326, p = .584. partial η^2 = .039. The results of this analysis, suggests that the order of task complexity did not affect the performance for either task complexity demand. Starting with the sunny or starting with the rainy condition did not make a difference in terms of performance.

Then, to respond the third research question that is: *Is there a significant progress of participants' performance rate between the first and second trial of task complexity?*, the test of within-subjects effects of the ANOVA repeated measures revealed that there was a significant difference of performance between the first and the second trial, F = 9.88, p = .014, partial η^2 = .553. That means that there was a significant progress between the first and second trial starting either conditions sunny or rainy conditions.

Overall, these results indicate that task complexity and order variation did not have a significant effect on task performance. First, the analysis revealed that there were no significant differences in rates of task performance when participants are assigned to perform under a low and high task complexity. Second, the analysis also showed that there were no significant differences in task performance between the training group that execute in the order sunny-rainy condition and the group that executed vice versa. Third, there was a significant performance progress between the first and second trial when the participants started with either of the conditions. Lastly, although the results revealed no significant differences in performance, a better performance was obtained when participants started with the sunny condition in both trials as Figure 6 shows.

Figure 6

Estimated Marginal Means of performance rates under the different orders.



The moderation effect of mental workload on task performance

To test the effects of mental workload on task performance that corresponds to the fourth research question: *Does mental workload have a moderation effect on task performance of participants when they are assigned to perform under different scenarios of task complexity?*, two Repeated Measures ANCOVA analysis were conducted in which the objective and subjective measures of mental workload were included as covariates, the variables of performance first and second trial as within-subjects factor, and the order as between-subject factor (order 0= start with sunny and order 1=start with rainy).

The first ANCOVA analysis shows a non-significant main effect of mental workload on task measures (F = 2.278, p = .175, η^2= .246). The interaction between performance and mental workload was non-significant (F = .363, p = .566, η^2= .049) as the test of within-subjects effects showed (Table 2). These results indicates that objective mental workload did not moderate the relation between task complexity and performance.

Table 2

Test of Within-Subjects Effects including the HR mean as Covariate.

Source		Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Performance * Mental Workload	Sphericity Assumed	136.59	1	136.59	.36	.57	.049
	Greenhouse-Geisser	136.59	1.00	136.59	.36	.57	.049
	Huynh-Feldt	136.59	1.00	136.59	.36	.57	.049
	Lower-bound	136.59	1.00	136.59	.36	.57	.049
Performance * Order	Sphericity Assumed	240.29	1	240.29	.64	.45	.084
	Greenhouse-Geisser	240.29	1.00	240.29	.64	.45	.084
	Huynh-Feldt	240.29	1.00	240.29	.64	.45	.084
	Lower-bound	240.29	1.00	240.29	.64	.45	.084

Likewise, the second ANCOVA analysis showed a non-significant main effect of subjective mental workload on task measures (F = 1.527, p = .256, η^2= .179). These results indicates that subjective mental workload did not moderate the relation between task complexity and performance. As well as the interaction between performance and subjective mental workload was no significant as the test of within-subjects effects showed (F = 3.376, p = .109, η^2= .325 (Table 3). Therefore, it might be concluded that neither objective nor subjective mental workload were not significantly related to the participant's performance.

Table 3

Test of Within-Subjects effects including the subjective MWL as Covariate.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Performance * Subjective MWL	Sphericity Assumed	901.73	1	901.73	3.38	.11	.32
	Greenhouse-Geisser	901.73	1.00	901.73	3.38	.11	.32
	Huynh-Feldt	901.73	1.00	901.73	3.38	.11	.32
	Lower-bound	901.73	1.00	901.73	3.38	.11	.32
Performance * Order	Sphericity Assumed	122.31	1	122.31	.46	.52	.06
	Greenhouse-Geisser	122.31	1.00	122.31	.46	.52	.06
	Huynh-Feldt	122.31	1.00	122.31	.46	.52	.06
	Lower-bound	122.31	1.00	122.31	.46	.52	.06

Discussion

The effect of task complexity and mental workload on task performance

The main purpose of this study was to test the effects of task complexity and mental workload on users' task performance between two training groups where the order of task complexity was varied. First, it was expected that the introduction of difficulties contained in the high task complexity produced higher task performance results in comparison with the low task complexity. It was hypothesized that there were significant differences among training groups in terms of performance results. However, the results of the ANOVA Repeated Measures revealed the contrary, there were not significant differences in terms of performance between the training group that execute under the sunny and the rainy conditions.

A possible explanation for these results may be due the lack of adequate levels of complexity in the structure of the task. The closest similarities between the sunny scenario and the rainy scenario might have contributed users to perceive similar levels of difficulty in both tasks, leaving aside the type of scenario. These results are consistent with research made by Mangos and Steele-Johnson (2001) who suggest that task difficulty is a factor that directly affects task performance. The difficulty of a task depends on the extent that individual perceive how demanding a task is (Campbell, 1988). Participants could have interpreted the two training scenarios with similar levels of difficulty. The acts executed during the first trial might have caused an overlap effect among the demands imposed by the second trial producing a redundancy effect. According to Wood (1986), the redundancy effect occurs when learning is hindered because additional information is presented compared to the presentation of less information. The redundancy effect might occur when identical information is given in two or more forms, the information is redundant and eliminating it might enhance learning. Some participants who did first the rainy conditions hampered their performance during second trial. While some participants who did first the sunny conditions perform better the second time. Thus, the same main compacting task executed during both trials and the perception of difficulty from the participants could have contributed to obtain not significant differences in performance between the high and the low task complexity.

Secondly, it was expected that there were significant differences in performance rates among the training groups under different order of task complexity. It was hypothesized that there were differences between the training group that first encountered the high task complexity and the group that first executed the low task complexity. However, the results of the ANOVA repeated measures showed that there was a not significant interaction between the order and performance results. Starting with sunny or with rainy weather conditions does not make difference in terms of performance. These results are contrary to those supported by the desired difficulty framework literature, which argues that the introduction of variations and or unpredictable events in the training environment causes difficulty for the learner but enhances performance (Bjork, 1994; Del Rey et al., 1982; Goettl, 1994; Shea & Morgan, 1979; Hall, Dominguez & Cavazos, 1994; Young, Cohen & Husak, 1993). In this case, the introduction of difficulties in form of rainy weather conditions did not enhance the participants' performance results due to the differences among training groups were not significant. One explanation of these effects might be because participants were novice learners since they did not have previous experience in the compacting process. According to Guadagnoli (2004), the performance of an individual might be predictable at any skill level. For example, for novice learners the performance outcome is expected to be high under easy conditions, as the task becomes more difficult, the expected level of performance drops rapidly. Thus, independently of which condition the participants made, they did not have experience in the compaction process, and they might assimilate both scenes as new ones and therefore leading to non-difference among training groups.

In addition, if it is considered the fact that the scenes were similar and did not produce the desired effect, this could contribute to have no significant difference in performance results. This effect is supported by Merbah & Meulemans, 2011 who argues that at the complexity of the task and the experience of the leaner determine the presence of contextual interference effect and complexity is not only dependent but also skill dependent (Merbah & Meulemans, 2011). Thus, the inexperience of the compaction process and the learners' perception of difficulty might explain the results.

The third expectation of this study was to find a significant progress between the first and the second trial independently whether the training groups encountered first sunny or rainy weather conditions. The results of the ANOVA Repeated Measures showed that there was a significant improvement in terms of performance for both training groups. This finding might be explained by the IBLT theory that argues that individuals learn by experience. Although half of participants did not have previous experience with VR, all the participants executed the main task two times. Participants could have used the first trial to learn and then apply their acquired knowledge in the second trial. This is in line with Gonzalez and Dutt (2011), who argue that the decision-maker interprets the situation, identifies a target state, and selects the appropriate action based on their experience. Participants could have built knowledge and experience from the first trial and then they could apply their new heuristics to perform the second time. Besides, the other half of the participants reported to have experience using VR previously. These participants could have taken this aspect in their favor, improving their performances rates during the second trial. These results are consistent with Jenkins et al. (2011), who argue that individuals based their decision-making process on their heuristics. That is the use of their knowledge and experience in a particular domain to make decisions. These knowledge and experience can be taught as a set of procedures by using information or demonstrations to form the basis of learning (Dreyfus, 1997). Thus, the experience from the first trial could be served as a basis to execute better during the second trial.

Although the ANOVA repeated measures analysis revealed a non-significant difference in rates of performance among the two training groups, the estimated marginal means revealed an interesting finding. The results showed that the participants who execute the task under the order sunny-rainy slightly obtained better performance scores in comparison to the group with the order rainy-sunny conditions. In this case, the contextual interference caused by the high task complexity did not produce significant differences in the performance results among groups, but slightly better performance results were obtained when participants started with the sunny condition. Therefore, the results support the traditional approach which argues that executing first the simple task and then the complex task to obtain better performance outcomes. The mechanism behind the traditional approach indicated that skill acquisition in novice subjects tends to be higher in low interference conditions (Del Rey, 1982; Hall, Dominguez & Cavazos, 1994; Shea, Kohl, & Indermill, 1990).Thus, it might be concluded that the introduction of variated task complexity did not enhance the performance results as it was expected but a slightly better result was obtained when participants first execute in sunny conditions.

The fourth research question was to test the modulator effect of mental workload on task performance for both training groups varying the order of task complexity. The results of the ANCOVA analysis showed that mental workload did not have a moderator effect on task performance. The results of the analysis indicated that the interaction between mental workload and performance was not significant. Thus, mental workload did not moderate the performance. This effect might be explained by a little stimulation of the physiological measures to predict performance. The rainy weather condition could be not demanding enough, as it was expected to cause changes in heart rate measures. Thus, similar heart rate measures in sunny and rainy conditions might produce an effect of insensitivity of the physiological measures and a poor effect of the mental workload to predict performance. The effect of heart rate insensitivity on task performance has been reported by previous research Mansika (2016), Shakouri (2018) and Sakib (2020). The authors have reported a similar effect in performance rates when training groups are exposed to perform different task complexities and no significant changes in the heart measures were found. These results are consistent with the underload effect described by Young et al., (2015) who explains that when an imposed demand causes little stimulation, an underload effect can be produced which is a consequence of a low level of engagement in the task. The rainy weather condition could be considered by the participants equally demanding as the sunny condition, resulting in similar reactions to respond to the demands in the two scenarios.

Another explanation of this effect might be due to the acclimatization effect. Young et al. (2015) also explain that the underload effect is compensated by the investment of additional resources which results in increased mental workload but can lead to a positive adaptation. Participants were exposed two times to the simulated environment, and between each change of scenario, participants did a short break in which they were interrupted to save their performance scores. This action could have produced an acclimatization effect. According to Stuiver et al. (2014), when there is an increase of mental demands, cardiovascular activity might respond in two ways: an increase in heart rate (initial reaction) or in a decrease of heart rate (regulation effect). The initial reaction of the VR application and the regulation effect during the execution of the second trial could have combined producing similar heart rate measures. As a result, the regulation effect could have contributed to perform better the second time compensating the performance scores achieved during the first trial. Besides, during first trial participants could get some experience and could modify their strategy to cope with the goal during the second trial. This is in line with Waard (1996) research that states that the additional effort investment by the operator depends on internal goals and strategies which depends on the structure of the task, the amount of practice and experience and the operator’s state (Waard, 1996). The effort expanded during the second trial could explain the similar performance outcomes from the two training scenarios.

In addition, other individual factors might explain the results. Although, psychophysiological measures such as heart rate have the advantage to detect mental workload continuously and unobtrusively, the individual variability is still a major challenge for many models. Objective mental workload is not a simple construct and is difficult to measure because it varies from person to person. In this project there were many variations that might be contributors of such effect. For instance, there were many differences in nationality, age, experience to drive and use VR technology, as well as the non-experience in compaction process. According to Hart & Staveland, (1988), the changes in heart rates could be attributable to other factors such as the physical condition of the user or emotional states which might affect task performance.

Theoretical implications

Results of this study broadly supports the work of other studies that confirms immersive VR as an effective training tool in construction industry (e.i. Cheng & Teizer, 2013; Goulding et al., 2012; Juang, Hung, & Kang, 2013; Joshi et al., 2020; Li, Chan & Skitmore, 2012; Rezazadeh et al., 2011; Sacks, Perlman & Barak, 2013; Vahdatikhaki et al., 2019). As mentioned in the literature review, only a few limited studies have included the constructs of task complexity, mental workload and task performance, involving the use of physiological measures, and those that have included them are conducted in the military field (i.e. Sakib et al., 2020; Shakouri et al. 2018) or in engineering operations (i.e. Chao et al., 2017; Das, Maiti and Krishna, 2020). The present study forms part of this limited research, innovating the approach to test the moderator effect of mental workload on task performance. The present might be considered as one of the first studies that evaluate the effects of variated order of task complexity and mental workload on task performance in the asphalting construction industry.

This study contributes to the understanding of the effect that produces mental workload on task performance by comparing two training groups executing under different order of task complexity. The results not only showed that task complexity and mental workload do not have influence on the performance results, but so it was also demonstrated that the levels of mental workload caused by the VR system are suitable for users to perform adequately. The results suggest that the mental workload caused by the VR session is within the stablished parameters according to the model of Young et al. (2015). Overall, this experimental study suggests that the order of task complexity in a varied order do not cause overload effects for the user using an immersive VR training. Therefore, this finding provides a basis support for developing future VR training systems in the asphalting construction industry. For instance, more interactive elements might be added to enhance the training experience for the learners and enhance performance. These results also may help practitioners to understand the mechanism of task complexity and mental workload on task performance in immersive VR to adopt it as educational tool.

Practical implications

Findings of the present research are expected to add value to the rapidly expanding field of mental workload prediction systems based on physiological measures, inspiring the design of a personalized training system for construction industry workers. The current study utilized a multi approach to measure mental workload and examined its impact on task performance of users. Practitioners could use the results of this study as a basis to create a personalized training systems for the asphalting construction industry. The design of this study may be applied in an adopted VR training system, by including more elements of interactivity through for example the introduction of more feedback clues to support learners to achieve the optimal performance.

Moreover, the proposed training approach could help to introduce new employees to practice some basic operations of the compacting process. The simulated environment and the dimensions of the vehicle are important elements that the proposed VR training application contained, and these can be used to train the spatial and motor skills for learners. This could help practitioners to induce the learners for familiarization of the scenario and machinery dimensions before to practice in the real setting.

Limitations

The generalizability of these results is subject to certain limitations. First, the scope of this study was limited to fourteen participants but only ten of them completed all the training sessions. Two participants presented symptoms of cybersickness, and the physiological data of the other two participants was not saved correctly. Thus, the small sample was reduced to ten participants, which might limit the significance of this study. Due to the current pandemic of covid-19 there was not possible to get more participants. These reduced sample might be not able to be representative enough. Notwithstanding the relatively limited sample, this work offers valuable insights into the extent of how mental workload might influence on participants' performance results.

The second limitation lies in the fact that not all the statistical assumptions were met. Two out four assumptions for running the ANCOVA Repeated Measures Analysis were violated. These assumptions were the linear relationship between the covariate (mental workload) and dependent variables (performance) and thus the assumption of homogeneity regression slopes. Although there was a partial linear relationship between mental workload and performance results in trial one, there was not a same condition for trial two. The scatter plots showed a linear relationship between the mental workload and the performance results in trial one, but in the second trial the relationship is crossed. Thus, the overall regression model can be inaccurate. According to Field (2016), if the relationship between the dependent variable and the covariate differs across the groups, then the overall regression might be inaccurate because it does not represent all the groups. In this study, the assumption of homogeneity of regression slopes and linear relationship between the mental workload and performance were violated. Although the researcher constructed the two-trial scatterplot for testing homogeneity of regression slopes for the two trials, note that a partial linear relationship was obtained. Thus, the results of the current study should be interpreted with some caution.

The third limitation was related to the physiological measures of mental workload. Although, psychophysiological measures such as heart rate have the advantage to detect mental workload continuously and unobtrusively, the individual variability is still a major challenge for many models. Mental workload is not a simple construct and is difficult to measure because varies from person to person. The changes in heart rates could be attributable to other factors such as the physical condition of the user or emotional states which might affect task performance (Hart & Staveland, 1988). External situations such as the current pandemic might be influenced the emotional state of the participants during the execution of the experiment leading them to get lower performance scores. Others could have been taking advantage of their emotional state to excel better. For instance, three participants experienced high levels of mental workload, but their levels of performance were good. Thus, physiological signals cannot only be attributed to task complexity but so physical characteristics, experience, emotional states, and levels of stress from the participants. Therefore, this study should not be considered as definitive.

However, this study maintains its value because it offers some insights into the levels of mental workload experienced by the participants during the execution of the two task complexities. The results revealed that the imposed task demands contained in the VR training did not cause excessive mental workload. The use of physiological measures allowed the research to obtain as much as could be possible an accurate estimation of mental workload for each participant during each trial. These measures were supported by survey instruments but many other physiological measures such as electrodermal and brain activity can be considered to support such effect.

Future research

This project was an experimental study aimed to investigate the effect of mental workload on task performance under different levels and order of task complexity using immersive VR to train the compaction process by using a simulated roller machine. The use of an immersive VR training approach demonstrated to be a practical cost-effective training method to provide users with important insights about the compaction process in the asphalting construction industry operation.

The current model is relevant to construct the basis for a more adaptative training models based on spatial and motor skills, taking advantage of the attributes of immersion that VR offer over other training methods. For instance, the introduction of variety of motor and problem-solving tasks in the training environment has demonstrated its effects to transfer training in real settings as Bjork, (1994) argues. Thus, to learn better, the learner must be stimulated via tactile, visual and auditory stimulus (Din et al., 1999). Such types of interaction might be adapted to the current model to create a more interactive VR system so that the learner could apply the knowledge from the VR to the real settings. The use of interactions generating real time actions might be interpreted and coordinated procedurally by the user, augmenting their learning though experience (Psotka, 1995). Interactive training environments have been demonstrated to provide means to transfer and accelerate training (Jarvis & Freitas, 2009). Thus, for example the use of spatial abilities to manipulate objects and feedback clues might be included to guide the learner during the whole training practice.

Secondly, the insights gained from this study might serve as a basis to test the transfer of training in real settings. Many studies have demonstrated positive learning transfer using immersive VR and then evaluate its effects in real settings. (Lackey et al., 2016; Luong et al., 2020; Mansikka et al., 2016; Perry et al., 2008; Sakib et al., 2020). Moreover, the use of 3D scenarios to train operational skills has demonstrated its advantages over 2D environments to transfer training (Rose et al., 2000). The 3D visualizations, the interaction with the elements, and the attributes of immersivity are some of the reasons demonstrated to have a better transfer of learning (Psotka, 1995). A further study could assess the long-term effects of the VR training in the real task. For instance, this version might be used as an introductory phase of training to get familiarized with the work environment and then assess the knowledge obtained when learners execute the task in the real setting to test its effects.

Thirdly, the insights gained from this study may contribute to establishing a methodological approach to estimate mental workload based on physiological measures for the construction industry. The latter can be used to design an adaptative workload predictive model based in physiological measures to enhance performance results. Modifications to the current VR training may be focalized to set adaptations to research the full picture of mental workload to enhance performance. For instance, the unobtrusive instruments to measure mental workload might form part of the current VR training systems to ensure safety for the learner and at the same time evaluate its impact during training to obtain the greatest benefits of an immersive VR.

Conclusion

Based on these interpretations it can be concluded that neither the order of task complexity nor mental workload has a significant impact in terms of performance. There are not significant differences in terms of performance when users executed under a low and a high task complexity, as well as when the order is varied (first sunny then rainy and vice versa). Executing under sunny first or rainy conditions did not make a significant difference in performance. However, there was a significant progress between the first and the second trial, independently of which condition the participants started. Although there were not significant differences in performance among training groups, slightly better performance results were obtained when participants started with the sunny condition. Despite its exploratory nature to test the effect of mental workload on task performance, this study offers acceptable levels of mental workload balance between the imposed tasks and the performance results, since the results showed low levels of mental workload during the practice. In overall, this training project might be used as a basis to develop as an effective training tool to be implemented in the asphalting construction industry.

References

Abbe, O. O., Harvey, C.M., Ikuma, L. H., & Aghazadeh, F. (2011). Modeling the relationship between occupational stressors, psychosocial/ physical symptoms and injuries in the construction industry. *International Journal of Industrial Ergonomics*, 41(2),106–117. https://doi.org/10.1016/j.ergon.2010.12 .002

Bornstein, M. H., Jager, J., & Putnick, D.L. (2013). *Sampling in developmental science: Situations, shortcomings, solutions, and standards,* 33 (4), 357- 370. https://doi.org/10.1016/j.dr.2013.08.003

Adhikarla, V.K., Wozniak, P., Barsi, A., Singhal, D., Kovács, P.T., & Balogh, T. (2014). Freehand interaction with large-scale 3D map data. *Proceedings of the 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE. https://ieeexplore.ieee.org/document/6874711

Agarwal, R., & Karahanna, E. (2000). *Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage* (pp. 665-694). MIS Quarterly. https://doi.org/10.2307/3250951

Asare, S. K., & McDaniel, L.S. (1996). The effects of familiarity with the preparer and task complexity on the effectiveness of the audit review process. *The Accounting Review,* 71(2), 139–159. https://www.jstor.org/stable/248443

Battig, W. F. (1979). The flexibility of human memory. In. L. S. Cermak & F. I. M. Craik (Eds.). *Levels of Processing in Human Memory* (pp. 33-44). Psychology Press. http://scholar.google.com/scholar_lookup?&title=The%20flexibility%20of%20human%20memory&pages=23-44&publication_year=1979&author=Battig%2CW.%20F.

Bjork, R. A. (1994). *Memory and metamemory considerations in the training of human beings* (Metcalfe J., & Shimamura Ed.). The MIT Press. https://www.gwern.net/docs/spaced-repetition/1994-bjork.pdf

Cain, B. (2007). *A review of the mental workload literature.* Defence research and development. Human integration section.

https://www.researchgate.net/publication/235159082_A_Review_of_the_Mental_Workload_Literature

Campbell, D. (1988). Task complexity: A review and analysis. *Academy of Management Review*, 13(1), 40-52. https://doi.org/10.5465/amr.1988.4306775

Carswell, C.M., Clarke, D., & Seales, W. B. (2005). Assessing mental workload during laparoscopic surgery. *Surgical innovation*, 12(1), 80–90. https://doi.org/10.1007/s00464-009-0522-3

Casner, S.M., Gore, B.F. (2010). *Measuring and evaluating workload: A primer.* NASA Technical Memorandum, (NASA/TM- 2010-216395). https://matbfiles.larc.nasa.gov/Workload_Primer_TM_Final.pdf

Chao, C.J., Sheng, Y.W., Yi-Jan, Y., Weng-Yan, F., & Feng-Yi, T. (2017). 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. https://doi.org/10.1002/hfm.20702

Cheng, T., & Teizer, J. (2013). Automation in construction real-time resource location data collection and visualization technology for construction safety and activity monitoring applications. *Automation in Construction*, 34(1), 3–15. https://doi.org/10.1016/j.autcon.2012.10.017

Csikszentmihalyi, M. (1996). *Where is creativity? In creativity: Flow and the psychology of discovery and invention* (pp. 23–50). New York: Harper Collins.

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. https://doi.org/10.1080/10508410903187547

Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, 41(1), 10-33. https://doi.org/10.1111/j.1467-8535.2009. 01038.x

Das, S., Maiti, J., & Krishna, O. B. (2020). Assessing mental workload in virtual reality based EOT crane operations: A multi-measure approach. *International Journal of Industrial Ergonomics*, 80(1), 2-13. https://doi.org/10.1016/j.ergon.2020.103017

Dede, C., Jacobson, J., & Richards, J. (2017). Introduction: Virtual, augmented, and mixed realities in education. In Liu D., Dede C., Huang R., & Richards, J., (Eds.). *Virtual, augmented, and mixed realities in education* (pp. 1-18). Springer. https://www.springer.com/gp/book/9789811054891

Del Rey, P., Wughalter, E., Dubois, D., & Carnes, M. M. (1982). Effects of contextual interference and retention intervals on transfer. *SAGE Journals, perceptual and motor skills*, 54(2), 467–476. https://doi.org/10.2466%2Fpms.1982.54.2.467

Dreyfus, H.L. (1997). Intuitive, deliberative and calculative models of expert performance. In Zsambok, C.E., Klein, G. (Eds.), *Naturalistic decision-making* (pp. 17-28). Lawrence Erlbaum Associates Mahwah NJ. https://doi.org/10.1002/9780470752937.ch15

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behaviour. *Psychological Review*, 66(3), 183-201. https://psycnet.apa.org/doi/10.1037/h0047707

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors*, 4(2), 59-74. https://doi.org/10.1177%2F001872086200400201

Empatica. (2020). Empatica E4 wristband. Empatica S.r.l. https://empatica.app.box.com/v/E4-User-Manual

European Construction Sector Observatory (2018, June). Country profile Netherlands. European Commission. https://ec.europa.eu › translations › renditions › pdf

Fan, J., Wade, J.W., Key, A.P., Warren, Z.E., & Sarkar, N. (2018). EEG-Based affect and workload recognition in a virtual driving environment for ASD intervention. *IEEE Transactions Biomedical Engineering*, 65(1), 43-51. http://dx.doi.org/10.1109/TBME.2017.2693157

Ferrer, R., & Dalmau, I. (2004). Revisión del concepto de carga mental: Evaluación, consecuencias y proceso de normalización. *Anuario de Psicologia*, 35(4), 521–545. https://revistes.ub.edu/index.php/Anuario-psicologia/article/view/8637

Field, A. (2016). *Discovering statistics using SPSS statistics (5th ed.).* SAGE Publications.

Goettl, B. P. (1994). Contextual interference effects on acquisition and transfer of a complex motor task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 38(18), 1220–1224. https://doi.org/10.1177%2F154193129403801817

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 118(4), 523–551. https://doi.org/10.1037/a0024558

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive science*, 27(4), 591–635. https://doi.org/10.1016/S0364-0213(03)00031-4

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufmen, & J. Thomas (Eds.), *Handbook of perception and human performance* (Chapter 41 pp. 41-49). New York Wiley. http://usd-apps.usd.edu/coglab/schieber/psyc792/workload/GopherDonchin1986.pdf

Goulding, J., Nadim, W., Petridis, P., & Alshawi, M. (2012). Construction industry offsite production: A virtual reality interactive training environment prototype. *Advanced Engineering Informatics*, 26(1), 103–116. https://doi.org/10.1016/j.aei.2011.09.004

Gutiérrez, M. J., Mora C.E., Añorbe-Díaz B., & González-Marrero A. (2017). Virtual technologies trends in education. *EURASIA Journal of Mathematics, Science and - Technology Education* 13(1), 470- 482. https://doi.org/10.12973/eurasia.2017.00626a

Hall, K. G., Domingues, E. & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. PubMed, 78(3), 835-841. https://www.krigolsonteaching.com/uploads/4/3/8/4/43848243/1994-hall.pdf

Hart, S.G., Staveland, L.E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey, J. A., & Picard, R. W. (2004). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6, 156-166. https://doi.org/10.1109/TITS.2005.848368

Heikoop, D. D., de Winter, J. C. F., van Arem, B., & Stanton, N. A. (2019). Acclimatizing to automation: Driver workload and stress during partially automated car following in real traffic. *Transportation Research Part F: Traffic Psychology and Behaviour*, 65, 503–517. https://doi.org/10.1016/j.trf.2019.07.024

Horswill, M.S., Marrington, S.A., McCullough, C.M., Wood, J., Pachana, N.A., McWilliam, J. & Raikos, M. K. (2008). The hazard perception ability of older drivers. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences,* 63(4), 212–18. https://doi.org/10.1093/geronb/63.4.P212

Jacko, J.A., & Ward, K.G. (1996). Toward establishing a link between psychomotor task complexity and human information processing. *Computers & Industrial Engineering* 31, 533–536. https://doi.org/10.1016/0360-8352(96)00192-1

Jarvis, S., & Freitas, S. (2009). Evaluation of an immersive learning programme to support triage training. *Proceedings of the First IEEE International Conference in Games and Virtual Worlds for Serious Applications,* 117-122. IEEE Computer Society. https://ieeexplore.ieee.org/document/5116561

Jebelli, H. (2019). *Wearable biosensors to understand construction workers' mental and physical stress* (Publication No. 0000-0003-4786-7616) [Doctoral dissertation, University of Michigan]. Deep blue repositories. https://hdl.handle.net/2027.42/149965

Jenkins, D. P., Stanton, N. A., Salmon, P. M., & Walker, G. H. (2011). A formative approach to developing synthetic environment fidelity requirements for decision-making training. *Applied Ergonomics,* 42(5), 757–769. https://doi.org/10.1016/j.apergo.2010.12.003

Joshi, S., Hamilton, M., Warren, R., Faucett, D., Tian, W., Wang, Y., & Ma, J. (2021). Implementing virtual reality technology for safety training in the precast/ prestressed concrete industry. *Applied Ergonomics*, 90, 2-10. https://doi.org/10.1016/j.apergo.2020.103286

Juang, J.R., Hung, W.H., & Kang, S.C. (2013). Sim Crane 3D+. *Advanced Engineering Informatics*, 27(4), 506–518. https://doi.org/10.1016/j.aei.2013.05.002

Lackey, S. J., Salcedo, J. N., Szalma, J. L., & Hancock, P. A. (2016). The stress and workload of virtual reality training: The effects of presence, immersion and flow. *Ergonomics*, 59(8), 1060–1072. https://doi.org/10.1080/00140139.2015.1122234

Leung, G.T.C., Yucel, G., & Duffy, V.G. (2010). The effects of virtual industrial training on mental workload during task performance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 20(6), 567–578. https://doi.org/10.1002/hfm.20198

Li, H., Chan, G., & Skitmore, M. (2012). Multiuser virtual safety training system for tower crane dismantlement. *Journal of Computing in Civil Engineering*, 26(5), 638–647. https://doi.org/10.1061/%28ASCE%29CP.1943-5487.0000170

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLoS ONE* 13(8), 1-30. https://doi.org/10.1371/journal.pone.0199661

Luong, T., Argelaguet, F., Martin, N., & Lecuyer, A. (2020). Introducing mental workload assessment for the design of virtual reality training scenarios. *Proceedings IEEE Conference on Virtual Reality and 3D User Interfaces*, 662-671. https://doi.org/10.1109/VR46266.2020.1581086856229

Magerko, B. & Laird, J. (2002). Towards building an interactive, scenario-based training simulator. *Proceedings of the 10th Computer Generated Forces and Behavior Representation Conference.* https://www.researchgate.net/publication/2890155_Towards_Building_an_Interactive_Scenario-based_Training_Simulator

Malone, T.W., & Lepper, M.R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In E.E. Snow & Farr (Eds.), *Aptitude, learning, and instruction*: *Conative and affective process analyses* (pp. 223-253). Lawrence Erlbaum Associates, Publishers. https://ocw.metu.edu.tr/pluginfile.php/2340/mod_resource/content/0/ceit706/week3/MakingLearningFun-ATaxonomyOfIntrinsicMotivationsForLearning.pdf

Mangos, P.M., & Steele-Johnson, D. (2001) The role of subjective task complexity in goal orientation, self-efficacy, and performance relations. *Human Performance*, 14(2), 169-186. https://psycnet.apa.org/doi/10.1207/S15327043HUP1402_03

Mansikka, H., Virtanen, K., Harris, D., & Simola, P. (2016). Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. *Applied Ergonomics*, 56, 213–219. https://doi.org/10.1016/j.apergo.2016.04.006

Marshall, T.E., & Byrd, T.A. (1998). Perceived task complexity as a criterion for information support. *Information & Management* 34(5), 251–263 https://www.sciencedirect.com/science/article/pii/S0378720698000573

Martin, D., Malpica, S., Gutiérrez, D., Masia, B., & Serrano, A. (2021). *Multimodality in VR: A survey.* Max Planck Institute for Informatics. https://www.researchgate.net/publication/348648114_Multimodality_in_VR_A_survey/link/6008f06192851c13fe26b8f8/download

Mascha, M.F. (2001). The Effect of task complexity and expert system type on the acquisition of procedural knowledge: Some new evidence. *International Journal of Accounting Information Systems* 2, 103–124. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.615.9243&rep=rep1&type=pdf

Merbah, S., & Meulemans, T. (2011). Learning a motor skill: Effects of blocked versus random practice a review. *Psychologica Belgica*, 51(1), 15-48. http://doi.org/10.5334/pb-51-1-15

Michaels, J., Chaumillon, R., Nguyen-Tri, D., Watanabe, D., Hirsch, P., Bellavance, F., Giraudet, G., Bernardin, D., & Faubert, J. (2017). Driving simulator scenarios and measures to faithfully evaluate risky driving behavior: A comparative study of different driver age groups. *PLoS ONE*, 12(10), 1–24. https://doi.org/10.1371/journal.pone.0185909

Mills, S., and Noyes J. (1999). Virtual reality: An overview of user-related design issues. *Interacting with Computers*, 11(4), 375–286. https://doi.org/10.1016/S0953-5438(98)00057-5

Moray, N. (Ed.). (2013). *Mental workload: Its theory and measurement* (Vol. 8). Springer Science & Business Media. https://www.springer.com/gp/book/9781475708868

Mulder, T., de Waard, D., & Brookhuis, K. A. (2005). Estimating mental effort using heart rate and heart rate variability. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendrick (Eds.), *Handbook of human factors and ergonomics methods* (pp. 1-8-20). CRC Press. https://research.rug.nl/en/publications/estimating-mental-effort-using-heart-rate-and-heart-rate-variabil-2

Ogden, G. D., Levine, M., & Eisner E. J. (1979). Measurement of workload by secondary tasks. *Human Factors,* 21(5), 529-548. https://doi.org/10.1177/001872087902100502

Peng L., & Zhizhong L. (2011). Toward understanding the relationship between task complexity and task performance. In Rau P.L.P. (Eds.), *Internationalization, Design and Global Development* (pp. 192-200). Springer Science + Business Media. https://doi.org/10.1007/978-3-642-21660-2_22

Pepinsky, P.N., Pepinsky, H.B., & Pavlik, W.B. (1960). The effects of task complexity and time pressure upon team productivity. *Journal of Applied Psychology*, 44(1), 34-38. https://psycnet.apa.org/doi/10.1037/h0039877

Perry, C. M., Sheik-Nainar, M. A., Segall, N., Ma, R., & Kaber, D. B. (2008). Effects of physical workload on cognitive task performance and situation awareness. *Theoretical Issues in Ergonomics Science,* 9(2), 95-113. https://doi.org/10.1080/14639220600959237

Psotka, J. (1995). Immersive training systems: Virtual reality and education and training. *Instructional Science, 23*(5-6), 405-431. https://doi.org/10.1007/BF00896880

Rao, A. K., Chahal, J. S., Chandra, S., & Dutt, V. (2019). Virtual-reality training under varying degrees of task difficulty in a complex search-and-shoot scenario. In Tiwary U., & Chaudhury S. (Eds.) *International Conference on Intelligent Human Computer Interaction* (pp. 248-258). Springer, Cham. https://doi.org/10.1007/978-3-030-44689-5_22

Rezazadeh, I. M., Wang, X., Firoozabadi, M., & Hashemi Golpayegani, M. R. (2011). Using affective human-machine interface to increase the operation performance in virtual construction crane training system: A novel approach. *Automation in Construction*, 20(3), 289-298. https://doi.org/10.1016/j.autcon.2010.10.005

Robinson, P. (2014). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22(1), 27-57. https://doi.org/10.1093/applin/22.1.27

Roscoe, A.H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *PubMed Biological psychology*, 34(3), 259–287. https://doi.org/10.1016/0301-0511(92)90018-P

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., & Penn, P. R. (2000). Training in virtual environments: Transfer to real world tasks and equivalence to real task training. *Ergonomics*, 43(4), 494-511. http://dx.doi.org/10.1080/001401300184378

Sacks, R., Perlman, A., & Barak, R. (2013). Construction safety training using immersive virtual reality. *Construction Management and Economics*, 31(9), 1005–1017. https://doi.org/10.1080/01446193.2013.828844

Sakib, M.N., Chaspari, T., Ahn, C.R., & Behzadan, A.H. (2020). An experimental study of wearable technology and immersive virtual reality for drone operator training. *Proceedings of the Workshop on Intelligent Computing in Engineering*, 154–163. EG-ICE. https://ris.utwente.nl/ws/portalfiles/portal/247003796/EG_ICE_2020.pdf

Salgado D., Roque F., Braga T., Keighrey C., Flynn R., Lázaro E., & Murray N. (2018). A QoE assessment method based on EDA, heart rate and EEG of a virtual reality assistive technology system. *Proceedings of the 9th ACM Multimedia Systems Conference*, 517–520. https://doi.org/10.1145/3204949.3208118

Schiessl, C. (2008). Subjective strain estimation depending on driving manoeuvres and traffic situation. IET Intelligent Transport Systems, 2(4), 258–265. https://doi.org/10.1049/iet-its:20080024

Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–218. https://doi.org/10.1111%2Fj.1467-9280.1992.tb00029.x

Schwarze, A., Kampling, H., Heger, O., & Niehaves, B. (2019). Is virtual reality the future of learning? A critical reflection. *Proceedings of the 52nd Hawaii International Conference on System Sciences,* 1759–1768. http://hdl.handle.net/10125/59616

Shakouri, M., Ikuma, L.H., Aghazadeh, F., & Nahmens, I. (2018). Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: The case of highway work zones. *International Journal of Industrial Ergonomics*, 66, 136–145. https://doi.org/10.1016/j.ergon.2018.02.015

Shea, C.H., Kohl, R., & Indermill, C. (1990). Contextual interference: Contributions of practice. *Acta Psychologica*, 73(2), 145-157. https://doi.org/10.1016/0001-6918(90)90076-R

Shea, J.B., & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*(2), 179–187. https://content.apa.org/doi/10.1037/0278-7393.5.2.179

Shi, Y., Zhu, Y., Mehta, R. K., & Du, J. (2020). A neurophysiological approach to assess training outcome under stress: A virtual reality experiment of industrial shutdown maintenance using Functional Near-Infrared Spectroscopy (fNIRS). *Advanced Engineering Informatics*, 46, 2-14. https://doi.org/10.1016/j.aei.2020.101153

Slater, M. (2003). A note on presence terminology. *Presence Connect*, 3(3), 1–5. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.3452&rep=rep1&type=pdf

Slater, M., & Sanchez-Vives M.V. (2016). Enhancing our lives with immersive virtual Rrality. *Frontiers in Robotics and AI*, 3, 1-74. https://doi.org/10.3389/frobt.2016.00074

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4), 75–76. https://doi.org/10.1111/j.1460-2466.1992.tb00812.x

Stuiver, A., Brookhuis, K.A., de Waard, D., Mulder, B. (2014). Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload. *International Journal of Psychophysiology*, 92(1), 35-41. https://doi.org/10.1016/j.ijpsycho.2014.01.010

Sumer, N. (2011) Cognitive and psychomotor correlates of hazard perception ability and risky driving. *Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* 211-216. https://drivingassessment.uiowa.edu/sites/drivingassessment.uiowa.edu/files/wysiwyg_uploads/031_sumer.pdf

Teizer, J., Cheng, T., & Fang, Y. (2013). Automation in construction location, tracking and data visualization technology to advance construction ironworkers' education and training in safety and productivity. *Automation in Construction*, 35, 53–68. https://doi.org/10.1016/j.autcon.2013.03.004

Tsang, P.S. & Vidulich, M.A. (2006). Mental workload and situation awareness. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 243-268). John Wiley & Sons, Inc. https://doi.org/10.1002/0470048204.ch9

Vahdatikhaki, F., Kassemi Langroodi, A., Makarov, D., & Miller, S. (2019). Context-realistic virtual reality-based training simulators for asphalt operations. *Proceedings of the 36th International Symposium on Automation and Robotics in Construction*, 218–225. https://doi.org/10.22260/isarc2019/0030

Waard, D. (1996). *The measurement of drivers' mental workload* (Publication No. 9068073087) [Doctoral dissertation, Groningen University]. Traffic Research Center Netherlands. https://research.rug.nl/en/publications/the-measurement-of-drivers-mental-workload

Wang, P., Wu, P., Wang, J., Chi, H. L., & Wang, X. (2018). A critical review of the use of virtual reality in construction engineering education and training. *International Journal of Environmental Research and Public Health,* 15(6). https://doi.org/10.3390/ijerph15061204

Wang, W. (2012). Workload assessment in human performance models using the secondary-task technique. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 56(1), 965–969. SAGE Publications. https://doi.org/10.1177%2F1071181312561202

Warwick, K., Gray, J., Roberts, D. (1993). An overview of virtual reality. *Virtual Reality in Engineering* (pp. 1-22). The Institution of Electrical Engineers. https://dl.acm.org/doi/10.5555/166665

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance*. Psychology Press. https://doi.org/10.4324/9781315665177

Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes, 37*(1), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

Xie, B., & Salvendy, G. (2000). Prediction of mental workload in single and multiple tasks environments. *International Journal of Cognitive Ergonomics*, 4(3), 213–242. https://doi.org/10.1207/S15327566IJCE0403_3

Young, D. E., Cohen, M.J. & Husak, W.S. (1993). Contextual interference and motor skill acquisition: On the process that influence retention. *Human Movement Science*, 12(5), 577-600. https://doi.org/10.1016/0167-9457(93)90005-A

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors*, 44 (3), 365– 375. https://doi.org/10.1518%2F0018720024497709

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, 58(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

Young, M. S., & Stanton, N. A. (2005). Mental workload. In N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. W. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods* (pp. 39–1-39-7). CRC Press. https://doi.org/10.1201/9780203489925

Zahabi, M., & Abdul R. A. M. (2020). Adaptive virtual reality-based training: A systematic literature review and framework. *Virtual Reality*, 24(4), 725–752. https://doi.org/10.1007/s10055-020-00434-w

Appendix A

Self-perception survey

Self-perception performance

1. What condition did you have to start with?
   a) Sunny condition
   b) Rainy condition

Instructions: Please mark with an "x" the scale that most represents your performance.

2. How do you rate your performance under sunny conditions?

1	2	3	4	5	6	7	8	9	10
Bad									Perfect

3. How do you rate your performance under rainy conditions?

1	2	3	4	5	6	7	8	9	10
Bad									Perfect

Self-perception mental workload

Instructions: Please mark with an "x" the scale that most represents your mental workload or stress.

1. In a scale of 1 to 5 please rate your level of stress experienced during this VR Training under sunny conditions

1	2	3	4	5
Not stressful at all	Moderately stressful	Considerable stressful	Very stressful	Extremely stressful

2. In a scale of 1 to 5 please rate your level of stress experienced during this VR Training under sunny conditions

1	2	3	4	5
Not stressful at all	Moderately stressful	Considerable stressful	Very stressful	Extremely stressful

Background information

Name: ____________________________________ Date: ___________________	
1. Age: a) 18-25 years b) 25-30 years c) 30-35 years d) 35-40 years	3. Experience driving a) 0-3 years b) 3-6 years c) 7-10 years
2. Gender: a) Male b) Female	4. How many times have you used Virtual reality technology? a) 0 times b) 1-3 times c) 4-6 times

Appendix B

Instructions to drive the roller machine.