Received: 12 July 2016 | Revised: 30 September 2016 | Accepted: 30 March 2017

DOI: 10.1002/hfm.20702

WILEY

# Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance

Chin-Jung Chao[1] | Sheng-Yu Wu[2] | Yi-Jan Yau[3] | Weng-Yan Feng[4] | Feng-Yi Tseng[5]

[1]Department of Industrial and System Engineering, Chung Yuan Christian University, Tao-Yum, Taiwan, R.O.C.

[2]Department of Industrial and System Engineering, Chung Yuan Christian University, Tao-Yum, Taiwan, R.O.C.

[3]System Development Center, National Chung-Shan Institute of Science and Technology, Tao-Yum, Taiwan, R.O.C.

[4]System Development Center, National Chung-Shan Institute of Science and Technology, Tao-Yum, Taiwan, R.O.C.

[5]System Development Center, National Chung-Shan Institute of Science and Technology, Tao-Yum, Taiwan, R.O.C.

**Correspondence**
Chin-Jung Chao, Department of Industrial and Systems Engineering, Chung Yuan Christian University, 200 Chung Pei Road, Chung Li District, Tao-yuan City, Taiwan 32023, R.O.C.
Email: davidchao@cycu.edu.tw

## Abstract

Industries will implement effective training programs to improve training performance, and an ideal training performance occurs under proper mental workload (MWL). Virtual reality (VR) has recently been widely utilized in training; however, only a few studies have investigated its effects on MWL and training performance simultaneously. The purpose of this study is to investigate the effects of VR training and traditional training methods, such as technical manuals (TM) and multimedia films (MF), on training performance and MWL. The results of the performance measurement show that VR training is considered the best training method compared to TM and MF, particularly in the case of complex tasks. The results of physiological measurements (GSR [galvanic skin response], LF% [low frequency], and LF/HF [high frequency] ratio) show a significant difference between reading TM and using computer (MF and VR), wherein the latter has a lower MWL. However, no significant difference in subjective MWL assessment (NASA-TLX [task load index]) and HF% measurement is found.

**KEYWORDS**

mental workload, NASA-TLX, performance, physiological indices, virtual reality

## 1 | INTRODUCTION

Training is very important for most businesses. In particular, high-risk and highly complex industries require cost-effective training programs to reduce training risks, to improve training performance, and to reduce training costs. Traditionally, the most commonly used training methods include the use of technical manuals (TM), multimedia films (MF), and practical training programs. However, these methods all have various limitations. For instance, operations cannot be viewed through various angles when using TM and MF, which are rendered in two dimensional (2D). Hence, the internal structure and operation processes of the machinery would be difficult to understand, and trainees can only rely on their imagination and memories. Practical training involves actual hands-on training but also comes with high training cost, risk, and logistic support problems. The internal structure and operation processes of machinery are also difficult to render with practical training.

In recent years, the rapid development of computer hardware and software, as well as cost reduction in hardware equipment, has raised the popularity in the use of three-dimensional (3D) virtual reality (VR) with computer simulations as a training method. Do et al. (2013) considered that one of the advantages of VR is to make an abstract concept concrete. Osberg (1995) and Bhagat, Liou, and Chang (2016) proved that VR could increase the motivation of users and cause them to focus more on learning. F. Lin, Ye, Duffy, and Su (2002) listed high interaction, less restrictive space, repeatability, flexibility, and low cost to be among the advantages of the VR training method. Duarte, Rebelo, and Wogalter (2010) have used VR to develop innovative technologies to improve man–machine interaction and to enhance training performance. Cates, Lönn, and Gallagher (2016) showed that using VR in training can improve performance by 17–49%. Trainees can become easily more familiar with machine operation when the VR training method is used. In manufacturing, trainees can view internal parts and assemblies of a machine clearly from various angles through VR. Moreover, VR not only presents more details of a machine's structure and operation procedures but also allows simulation of procedures for operation. Hence, trainees can fully understand the principles and procedures of operation. Grantcharov et al. (2004) and Lehmann et al. (2005) found that the doctors who have been trained using the VR system apparently had better-quality surgery skills. VR technology has also been widely utilized in trainings for design, manufacturing, transportation, military, and nuclear power plant (Chaffin, 2009; Chung, Shewchuk, & Williges, 2002; Duarte et al., 2010; Fuhua, Duffy, & Su, 2002; Hue,

  

Delannay, & Berland, 1997; Lee, Chou, & Sun, 2015; Nathanael, Mosialos, Vosniakos, & Tsagkas, 2016; Vilar, Rebelo, F., & Noriega, 2016; Wu, Mu, Yang, & Gu, 2012). Instead of the effects of VR on mental workload (MWL), applications of VR have often focused on training performance. Apart from increasing training performance, including enhancing familiarity and reducing error rate, risk, and time, the MWL of trainees should also be considered when creating a good training method (Gilbert, Leung, & Duffy, 2010). A training that requires too high or too low MWL is neither useful nor does it promote training performance. The relationship between training performance and MWL is similar to an inverted-U curve wherein the best training performance will occur under proper MWL (Hwang et al., 2008). Therefore, the influences of different training methods on MWL and training performance of operators are worth exploring.

A number of theories to measure MWL have been proposed (Cegrra & Chevalier, 2008; Charlton, 2002; Farmer & Brownson, 2003; Gilbert et al., 2010; Johnson & Widyanti, 2011; Lean & Shan, 2012). MWL measurement can be divided into work performance measurement, physiological measurement, and subjective rating. Researchers studying the work performance measurement aspect have indicated that work performance is restricted by limited mental resources and that major and minor work performances interact with each other (Dadashi, Stedmon, & Pridmore, 2013). Hence, measurement of the major or minor work performance can indicate the amount of MWL (Ogden et al., 1979; W. Wang, 2012). For the physiological measurement aspect, numerous studies have shown that the human physiological indices react from the various effects of psychological pressure (Charlton, 2002; Cobb, Nichols, Ramsey, & Wilson, 1999; Dahlstrom & Nahlinder, 2006; Fallahi, Motamedzade, Heidarimoghadam, Soltanian, & Miyake, 2016; Goldstein, Bentho, Park, & Sharabi, 2011; Lean & Shan, 2012; Reyes del Paso, Langewitz, Mulder, Roon, & Duschek, 2013; Ryu & Myung, 2005; Sztajzel, 2004; L. M. Wang, Duffy, & Du, 2007). Among the most widely used physiological indices are electroencephalogram, galvanic skin response (GSR), average heart rate (AHR), pupil sizes, and the relative parameters of heart rate variability (HRV), which include very-low-frequency (VLF) power, low-frequency (LF) power, high-frequency (HF) power, the ratio of LF to HF (LF/HF ratio), and LF and HF power in normalized units (LF% and HF%, respectively). The more severe the MWL, the higher the AHR, GSR, LF%, and LF/HF ratio but the lower the HF% and blink rate. In terms of individual psychological MWL indices, the NASA task load index (NASA-TLX) subjective questionnaire is one of the most widely used indices (Gilbert et al., 2010; Hart, 2006; Hart & Staveland, 1998). The questionnaire investigates workload when workers perform tasks through six subscales, including (i) mental demand, (ii) physical demand, (ii) temporal demand, (iv) overall performance, (v) effort, and (vi) frustration. The three former subscales are related to the MWL of operating a task, whereas the three latter subscales are related to the subjective experience of the subject. The higher scores of the integrated six dimensions are the higher MWL and vice versa (Rubio, Diaz, Martin, & Puente, 2004).

TM, MF, and VR training methods with two different task complexities: brake assemblies inspection and replacement (simple task) and fuel line subsystem inspection and parts replacement of a vehicle (complex task) were used as experimental variables to study their effects on MWL and training performance. The MWL of participants performing different training methods is compared with the NASA-TLX subjective questionnaire and objective physiological indices, such as AHR, GSR, and HRV-associated indices (LF%, HF%, and LF/HF ratio). The number of operation errors and operation time in maintenance tasks are used as performance indices to evaluate the effectiveness of training methods and task complexities.

## 2 | RESEARCH METHODS

### 2.1 | Subjects

Forty-eight students (24 men and 24 women) from Chung Yuan Christian University with ages ranging from 19 to 27 years participated in this experiment. They were randomly assigned to three groups to use TM, MF, or VR as training methods, and each subject had to execute both simple and complex maintenance tasks. Each group was composed of 16 subjects (8 men and 8 women), and each subject could use only the specified training method. All participants had normal eyesight after correction and had neither significant physical diseases nor operation experiences of the tasks used in the study. The subjects were directed not to take in food or drinks containing caffeine nor any medication before the day of the experiments. They were also asked to sleep for an average duration of seven hours before the day of the experiment.

### 2.2 | Apparatus

VR falls into three main categories, namely, nonimmersive (desktop), semi-immersive, and fully immersive systems (Mujber, Szecsi, & Hashmi, 2004). Using the VR system for training has many benefits; however, some associated side effects, such as cyber-sickness, pallor, sweating, dryness, fullness, vertigo, ataxia, disorientation, headache, eye strain, nausea, and vomiting also occur (LaViola, 2000). VR has many limitations, for example, field of view, graphical quality of the virtual environment, and interaction devices (J. W. Lin, Duh, Parker, Abi-Rached, & Furness, 2002; Ragan et al., 2015). These side effects are more likely to occur in a full immersive VR system and less likely to occur in a nonimmersive VR system (Robertson, Card, & Mackinlay, 1993). To eliminate these side effects, the nonimmersive VR system was provided as a training method in our experiment. The other two training methods included MF video and paper TM. The VR scenes were created by EON Studio, and the videos were made in the studio of Chung-Yuan Christian University and edited by EDIUS-6. The VR and MF trainings were performed on a desktop PC running on Windows 7 and an Intel-I5-3470 Quad core processor (3.20 GHz, 8GB RAM) with graphical card, NVIDIA GTX-950. The Kmplayer and EON viewer were installed to watch MF video and run VR. A 22-inch LCD screen with a resolution of 1920 × 1080 pixels and a Logitech Z130 speaker were used for scenario and audio output. A mouse and a keyboard were used to pause, to rewind, and to forward the video in the MF training, as well as to select and flip any object in the VR scene. Two prerecorded videos, two VR scenes, and two technical paper manuals were given

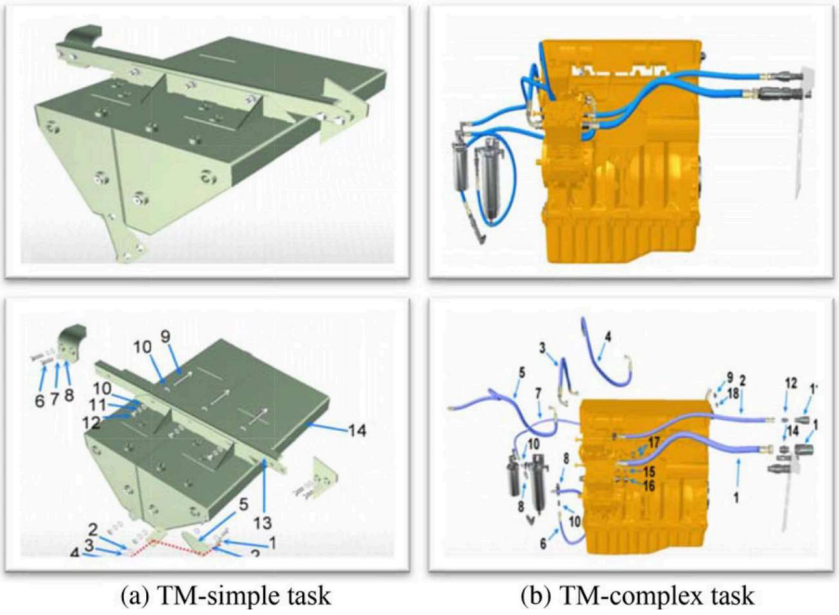


**FIGURE 1** Demonstration of TM tasks

for two separate TM, MF, and VR training phases in the simple and complex task sections. The length of each training phase was approximately 15 min. The VR scenes were sequential instructions that divided the tasks into multiple steps. In each step, subjects were not only allowed to move freely around the environment with the mouse or the keyboard but also interacted with each composed object to improve comprehension. The self-made brake and fuel line subsystems of a vehicle were used for maintenance tasks. A Power-Lab/16 sp system was connected to two amplifiers, namely, Dual-Bio-Amp and GSR-Amp. The former was utilized to acquire and amplify physiological signals, such as AHR and HRV-associated signals, whereas the latter was utilized to acquire and amplify GSR signals. The sensors of these apparatus were connected from the body-related parts to the Power-Lab/16 sp (the GSR sensors were hung on the fore and middle fingers of the nondominant hand). The transmission cables were long enough and placed on the nonworking path to avoid interference with the works. Chart analysis software was used to start and suspend physiological signal retrievals.

## 2.3 | Experimental procedures and variables

### 2.3.1 | Procedures

This experiment involved three stages, namely, (i) preexperiment, (ii) experimental execution, and (iii) questionnaire filling stages. At the preexperiment stage, the experimenters explained the experimental purpose, procedures, and all possible situations, for example, the syndromes of side effects that could happen to the subjects in the experiment. The subjects were requested to reply to the experimenters, and the experiment would be suspended if participants felt that they were affected by the side effects. The subjects completed a consent form as experimental participants and listened to light music to relax. After the experimental apparatuses were corrected, the experimenters assisted the subjects in wearing the sensors. The physiological signals-at-rest of the subjects were measured. The measurement was taken approximately 35 min prior to the start of the experimental execution stage. The experimental execution stage was divided into complex (lasted approximately 75 min) and simple (lasted approximately 50 min) task sections. Each section was classified into training and practical operation phases. During the training phase, the subjects randomly selected simple or complex training for approximately 15 min in any of the following: reading TM, watching MF, or using VR while their physiological signals were measured. The measurement started and suspended at the beginning and end of the two training phases, respectively. Once the training method was selected, the subjects can only use the selected method in the next training phase. The practical operation phase would not be executed until the subjects had completely understood the training contents and had confirmed readiness to the experimenters. The subjects rested for 5 min after the training phase and then were asked to execute the simple or complex task in which they had been trained. After finishing one task section, the subjects were asked to execute another section. The operation time and number of operation error of each subject were recorded when he or she executed the simple and complex tasks. After the experimental execution stage, the subjects were requested to take about 10 min to fill in a NASA-TLX questionnaire to determine their subjective MWL of the training method, which they had received during training phases. The experiment took approximately 170 min.

### 2.3.2 | Variables

#### Independent variables

Independent variables include training method and task complexity of vehicle maintenance operations. The former included three levels, namely, using TM, MF, or VR of vehicle maintenance operations as a medium for training subjects. The latter has two levels; namely, simple and complex vehicle maintenance operations (see Figures 1, 2, and 3, respectively). The properties of the simple and complex tasks were

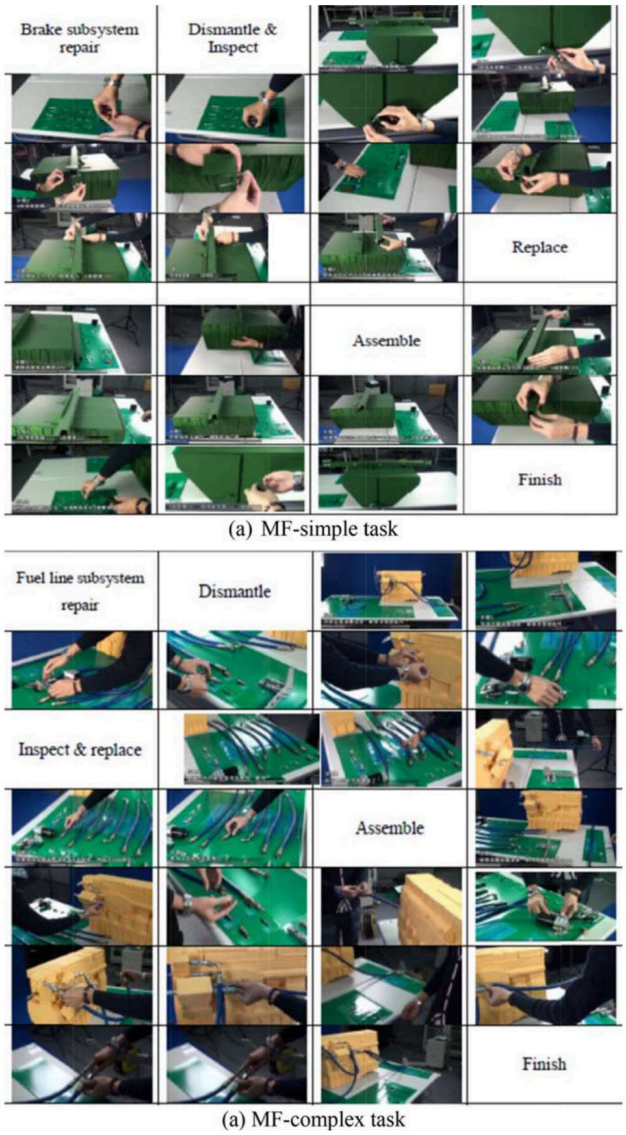


**FIGURE 2** Demonstration of MF tasks

different. The tasks at the simple level, which consists of 32 steps and requires 6 kinds of tools, involve tightening/loosening screws and replacing parts of a brake assembly, whereas the tasks at the complex level, which includes 37 steps and requires 6 kinds of tools, involve the inspection and replacement of the parts of a fuel line subsystem of a vehicle. A pretest was conducted to ensure that the complex levels of the two tasks were different. The subjects read the TM in front of the work platform, watched the MF, and interacted with the VR through a computer.

#### Dependent variables

AHR, GSR, and HRV-associated indices (LF%, HF%, LF/HF ratio), operation time, number of operation errors, and scores of NASA-TLX subjective questionnaire were collected as dependent variables.

### 2.3.3 | Analysis methods

A two-way mixed design analysis of variance (ANOVA) was used to examine the effects of training method and task complexity on the dependent variables. The interaction between training method and task complexity was also tested in addition to their main effect. A simple main effect would be tested if their interaction is significant; that is, the effect of training method on different levels of task complexity and the effect of task complexity on different training methods would be examined. When the interaction is not significant, only the main effects of training method and task complexity of maintenance tasks are examined. The critical differences for paired comparison are determined using by least significant difference (LSD) when the main effect is significant and the levels of independent variable are over or equal three levels. In terms of the subjective mental workload assessment, a one-way ANOVA was conducted to evaluate the scores of NASA-TLX questionnaire on mental workload. If the results were significant, LSD was used for post hoc comparison.

## 3 | EXPERIMENTATION RESULTS

### 3.1 | Training performance

#### 3.1.1 | Operation errors

The mean operation errors for the various training methods and task complexities are shown in Table 1. The ANOVA results indicate that both training method and task complexity have significant effects on mean operation errors, $F(2, 45) = 22.04$, $p < .05$; $F(1, 45) = 226.42$, $p < .05$. A significant interaction between these two factors is also observed, $F(2, 45) = 3.54$, $p < .05$. Hence, the simple main effects and LSD post hoc comparison of the training method factor was performed. Their results indicate that the mean operation errors of training by TM (4.13 times) are significantly ($p < .001$) higher than those of training by MF (1.56 times) and VR (1.38 times), but no significant ($p > .05$) difference is observed between training by MF or VR under simple tasks (TM > MF, VR). Under complex tasks, the mean operation errors of training by TM (9.75 times) are significantly ($p < .001$) higher than those of training by MF (7.19 times), which are also significantly ($p < .001$) higher than those of training by VR (5.13 times; TM > MF > VR). The operation errors of complex tasks are also all significantly ($p < .001$) higher than those of simple tasks under all kinds of training method.

#### 3.1.2 | Operation time

The mean operation time for the various training methods and task complexities is shown in Table 2. The ANOVA results indicate that both training method and task complexity have significant effects on mean operation time, $F(2, 45) = 87.01$, $p < .05$; $F(1, 45) = 720.13$, $p < .05$. Their interaction is also significant, $F(2, 45) = 9.33$, $p < .05$. Simple main effect and LSD post hoc comparison were performed, and their results indicate that the mean operation time of training by TM (865.38 s in simple task, 1383.19 s in complex task) is significantly ($p < .001$) longer than that of training by MF (673.81 s in simple task, 968.94 s in complex task), which is also significantly ($p < .001$) longer than that of training by VR (421.19 s in simple task, 715.00 s in complex task; TM > MF > VR) under any simple or complex tasks. The mean operation time of

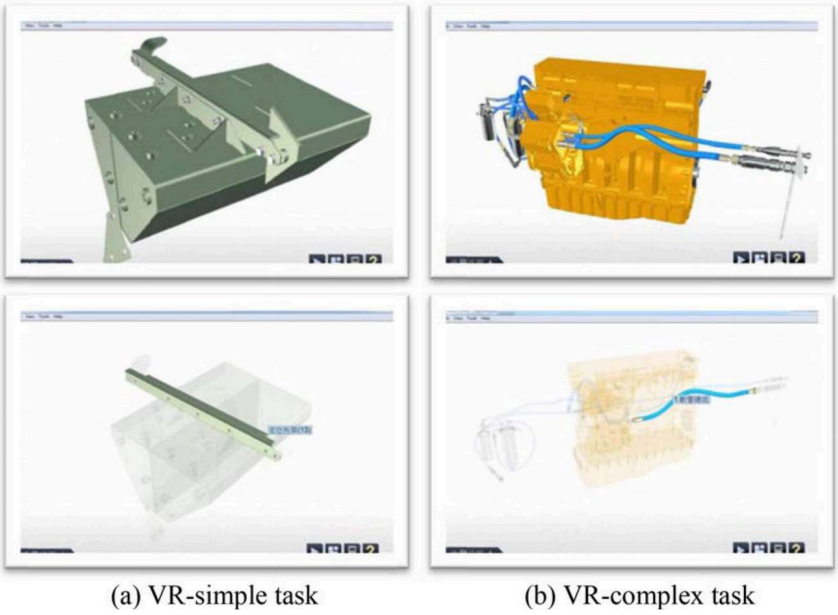


**FIGURE 3** Demonstration of VR tasks

**TABLE 1** Summary of descriptive statistics of the operation errors

| Training Method | Mean Operation Errors ± Standard Deviation (times) | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 1.38 ± 0.81 | 5.13 ± 1.54 | 3.25 ± 1.06 |
| MF | 1.56 ± 0.96 | 7.19 ± 2.46 | 4.38 ± 1.42 |
| TM | 4.13 ± 2.68 | 9.75 ± 2.52 | 6.94 ± 2.15 |
| Average | 2.35 ± 2.10 | 7.35 ± 2.89 | 4.85 ± 2.22 |

*Note:* MF = multimedia film; TM = technical manual; VR = virtual reality.

**TABLE 2** Summary of descriptive statistics of operation time

| Training Method | Mean Operation Times ± Standard Deviation (s) | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 421.19 ± 79.21 | 715.00 ± 121.11 | 568.09 ± 1.97 |
| MF | 673.81 ± 93.81 | 968.94 ± 78.53 | 821.38 ± 70.56 |
| TM | 865.38 ± 134.72 | 1383.19 ± 212.39 | 1124.28 ± 168.11 |
| Average | 653.46 ± 210.71 | 1022.38 ± 313.85 | 837.92 ± 257.77 |

*Note:* MF = multimedia film; TM = technical manual; VR = virtual reality.

**TABLE 3** Summary of descriptive statistics of GSR

| Training Method | Mean GSR Values ± Standard Deviation ($\mu$s) | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 2.44 ± 0.83 | 2.67 ± 0.88 | 2.56 ± 0.63 |
| MF | 2.59 ± 1.14 | 3.36 ± 1.89 | 2.98 ± 1.26 |
| TM | 3.31 ± 1.32 | 4.26 ± 1.78 | 3.79 ± 1.32 |
| Average | 2.78 ± 1.16 | 3.43 ± 1.68 | 3.11 ± 1.21 |

*Note:* GSR = galvanic skin response; MF = multimedia film; TM = technical manual; VR = virtual reality.

**TABLE 4** Summary of descriptive statistics of AHR

| Training Method | Mean AHR Values ± Standard Deviation (bpm) | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 81.04 ± 8.86 | 85.70 ± 6.97 | 83.37 ± 6.92 |
| MF | 84.41 ± 5.64 | 87.53 ± 6.82 | 85.97 ± 5.70 |
| TM | 94.37 ± 11.62 | 95.14 ± 11.40 | 94.76 ± 11.02 |
| Average | 86.61 ± 10.54 | 89.46 ± 9.43 | 88.03 ± 9.41 |

*Note:* AHR = average heart rate; MF = multimedia film; TM = technical manual; VR = virtual reality.

complex tasks is significantly ($p < .001$) longer than that of simple tasks regardless of the type of training method.

## 3.2 | Physiological responses

### 3.2.1 | Galvanic skin response

A lower measured GSR value suggests the person is more relaxed and vice versa. The mean GSR values for various training methods and task complexities are shown in Table 3. The results of ANOVA indicate that both training method and task complexity have significant effects on the GSR value, $F(2, 45) = 87.01$, $p < .05$; $F(1, 45) = 7.98$, $p < .05$, respectively, whereas their interaction is not significant, $F(2, 45) = 0.90$, $p > .05$. An LSD post hoc test was performed to compare the means of the subsets of training method. The results indicate that the mean GSR value of training by TM (3.79 $\mu$s) is significantly ($p < .05$) higher than that of training by MF (2.98 $\mu$s) and VR (2.56 $\mu$s), but there is no significance between MF and VR (TM > MF, VR). Regarding task complexity, the mean GSR value (3.43 $\mu$s) of the complex task is significantly ($p < .01$) higher than that of the simple task (2.78 $\mu$s).

### 3.2.2 | Average heart rate

A higher AHR value suggests the person has a heaver MWL and vice versa. The mean AHR values for various training methods and task complexities are shown in Table 4. The ANOVA results indicate that

**TABLE 5** Summary of descriptive statistics of LF%

| Training Method | Mean LF% Values ± Standard Deviation | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 47.37 ± 15.75 | 53.45 ± 19.05 | 50.41 ± 14.30 |
| MF | 50.78 ± 16.17 | 56.85 ± 18.62 | 53.81 ± 11.96 |
| TM | 60.41 ± 13.95 | 68.64 ± 13.94 | 64.52 ± 11.55 |
| Average | 52.85 ± 15.99 | 59.64 ± 18.21 | 56.25 ± 13.80 |

*Note:* LF = low frequency; MF = multimedia film; TM = technical manual; VR = virtual reality.

both training method and task complexity have significant effects on the AHR value, $F(2, 45) = 8.47$, $p < .05$; $F(1, 45) = 8.73$, $p < .05$, respectively, but no significant interaction, $F(2, 45) = 1.37$, $p > .05$, is observed between these factors. The main effects of training method and task complexity are significant and, thus, an LSD post hoc test was conducted to compare the means of the subsets. The results of the LSD post hoc test indicate that the mean AHR value of training by TM (94.76 bpm) is significantly ($p < .001$) higher than that of training by MF (85.97 bpm) and VR (83.37 bpm). Nevertheless, no significant ($p > .05$) difference is observed between training by MF and VR (TM > MF, VR). Regarding task complexity, the mean AHR value (89.46 bpm) of complex tasks is significantly ($p < .05$) higher than that of the simple tasks (86.61 bpm).

### 3.2.3 | Heart rate variability

In this study, HRV is based on the spectrum analysis calculated from the power of a specific frequency range. The LF% and HF% values are indices of sympathetic and parasympathetic nerve activities. The LF/HF ratio of the parasympathetic activity is an index of sympathetic/parasympathetic balance.

#### Low-frequency component

LF% controlled by both systolic and diastolic pressures from peripheral blood vessels can be regarded as an index of sympathetic nerve activity. The higher LF% of a person represents higher sympathetic nerve activity and refers to the increase in its intensity. Therefore, the LF% value can be regarded as an index of environmental pressure. A higher LF% value suggests the person has a heaver MWL, and vice versa. The mean LF% values for various training methods and task complexities are shown in Table 5. The effects of training method and task complexity on LF% value were analyzed using ANOVA, and the results indicate that training method and task complexity have significant effects on LF% value, $F(2, 45) = 5.42$, $p < .05$; $F(1, 45) = 5.14$, $p < 0.05$. No interaction between the two factors is observed, $F(2, 45) = 0.06$, $p > .05$. The main effects of training method and task complexity are significant and, hence, an LSD post hoc test was conducted to compare the means of the subsets. The results indicate that mean LF% value of training by TM (64.52) is significantly ($p < .05$) higher than that of training by MF (53.81) and VR (50.41). However, no significant difference between training by MF and VR is observed (TM > VR, MF). Regarding task complexity, the mean LF% value (59.64) of complex tasks is significantly higher than that of simple tasks (52.85).

**TABLE 6** Summary of descriptive statistics of HF%

| Training Method | Mean HF% Values ± Standard Deviation | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 27.22 ± 7.51 | 25.97 ± 2.82 | 26.60 ± 4.58 |
| MF | 26.57 ± 5.09 | 26.26 ± 5.47 | 26.41 ± 4.45 |
| TM | 25.48 ± 7.04 | 24.50 ± 9.40 | 24.99 ± 4.10 |
| Average | 26.42 ± 6.53 | 25.58 ± 6.40 | 26.00 ± 4.35 |

*Note:* HF = high frequency; MF = multimedia film; TM = technical manual; VR = virtual reality.

**TABLE 7** Summary of descriptive statistics of the LF/HF

| Training Method | Mean LF/HF Ratios ± Standard Deviation | | |
|---|---|---|---|
| | Simple | Complex | Average |
| VR | 1.88 ± 0.80 | 2.13 ± 0.92 | 2.00 ± 0.70 |
| MF | 1.98 ± 0.80 | 2.25 ± 0.90 | 2.12 ± 0.64 |
| TM | 2.68 ± 1.28 | 3.37 ± 1.92 | 3.02 ± 0.99 |
| Average | 2.18 ± 1.03 | 2.58 ± 1.42 | 2.38 ± 0.90 |

*Note:* HF = high frequency; LF = low frequency; MF = multimedia film; TM = technical manual; VR = virtual reality.

#### High-frequency component

HF% is related to respiration and can be regarded as a parasympathetic nerve activity index. A higher HF% value suggests that the person is more relaxed. On the contrary, the lower the HF% value, the more nervous or excited the person. As a result, HF% value can also be regarded as an index of environmental pressure. A lower HF% value suggests the person has a heaver MWL and vice versa. The mean HF% value for various training methods and task complexities are shown in Table 6. The results of ANOVA indicate that neither training method nor task complexity has a significant effect on the mean HF% value, $F(2, 45) = 0.64$, $p > .05$; $F(1, 45) = 0.36$, $p > .05$. Their interaction was not significant, $F(2, 45) = 0.04$, $p > .05$). Although training method and task complexity have no significant effect on HF%, the HF% of training by TM is lower than that of training by MF, whereas that of training by MF is lower than that of training by VR. This result is similar to the HF% of simple task being lower than that of complex task. From these results, we can infer that the mental workload of training by TM could be the largest, followed by MF and VR, and that the mental workload of complex tasks is higher than that of simple tasks.

#### Sympathetic/parasympathetic ratio

LF/HF ratio refers to the ratio between the power of LF and HF bands and can be regarded as the overall balance between sympathetic and parasympathetic systems. A higher LF/HF ratio implies the domination of the sympathetic system, whereas a lower one reflects the domination of the parasympathetic system. For a healthy adult at rest, the LF/HF ratio should be around 1 to 2 and increases as the tension of the person increases. As a result, LF/HF ratio is often used as an indicator of pressure being borne by a person. A higher value of LF/HF ratio suggests the person has a heaver MWL and vice versa. The mean values of the LF/HF ratio of the training method and task complexity are shown in Table 7. The results of ANOVA reveal that training method has a significant effect on the LF/HF ratio, $F(1, 45) = 7.96$, $p < .05$. However, task

**TABLE 8** Summary of NASA-TLX subjective evaluation on training method

| Title | Training Methods | Average | *F* | *P* |
|---|---|---|---|---|
| Mental demand | VR | 70.38 | 0.95 | .396 |
| | MF | 74.81 | | |
| | TM | 79.63 | | |
| Physical demand | VR | 48.06 | 0.66 | .523 |
| | MF | 41.50 | | |
| | TM | 37.06 | | |
| Temporal demand | VR | 50.25 | 2.89 | .066 |
| | MF | 44.25 | | |
| | TM | 63.81 | | |
| Performance | VR | 50.63 | 0.54 | .589 |
| | MF | 56.63 | | |
| | TM | 58.75 | | |
| Effort | VR | 53.56 | 0.68 | .514 |
| | MF | 62.13 | | |
| | TM | 56.50 | | |
| Frustration | VR | 61.13 | 0.68 | .510 |
| | MF | 52.44 | | |
| | TM | 58.88 | | |
| Overall | VR | 60.78 | 1.48 | .239 |
| | MF | 61.29 | | |
| | TM | 67.69 | | |

*Note:* MF = multimedia film; NASA-TLX = NASA-task load index; TM = technical manual; VR = virtual reality.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

complexity has no significant effect on the LF/HF ratio, $F(2, 45) = 0.31$, $p > .05$. Their interactions has no significant effect on the LF/HF ratio, $F(2, 45) = 0.31$, $p > .05$. The effects of training method on LF/HF ratio are significant, and thus, an LSD post hoc comparison was conducted to compare the differences in various training methods. The results indicate that the mean LF/HF ratio (3.02) of training by TM is significantly ($p < .001$) higher than that of training by MF (2.12), as well as VR (2.00). No difference between training by MF and VR is observed (TM > VR, MF). Generally, a trend was observed that when a complex task was carried out the LF/HF ratio was higher regardless of the type of training method. For both simple and complex tasks, training by VR would probably have the lowest mental workload among the three training methods.

### 3.3 | NASA-TLX subjective workload assessment

One-way ANOVA was used to analyze the scores of NASA-TLX questionnaire. Fisher's LSD method was used to compare the differences between the various training methods when the main effect is significant. The scores of NASA-TLX questionnaire and the results of ANOVA are listed in Table 8, which shows that no significant difference ($p > .05$) in the training methods is observed for each subscale and the overall workload. Although this result implies that training method has limited effect on MWL, training through VR yielded the lowest scores in the subscales of mental demand, performance, effort, and overall. However, training by VR has the highest scores in physical demand and frustration subscales.

## 4 | DISCUSSION

### 4.1 | Effects of training method and task complexity on training performance

The various training methods, task complexities, and their interaction have significant effects on training performance. Under simple tasks, training by TM resulted in greater operation errors than training by MF and VR, with the results of the latter two having no significant differences. Under complex tasks, the number of operation errors for training methods is significant. Training by TM also incurred the largest number of operation errors, followed by MF and VR, which is significantly smaller than MF. The operation errors of complex tasks are all larger than those of simple tasks regardless of the type of training method. In terms of operation time, the effects of training method, task complexity, and their interaction are all significant. Regardless of the kind of task complexity, the operation time of training by TM is found to be significantly longer than that of training by MF, which is also significantly longer than that of training by VR. The operation time of training in complex tasks is significantly longer than that of training in simple tasks regardless of the type of training method. Based on these results, VR is considered the best training method compared to TM and MF, particularly in the case of complex tasks. One of the reasons could be because VR systems are easier to make abstract concepts concrete and have better interaction between the subjects and training contents. The subjects could be more likely to understand the concepts behind the exteriors and adjust the learning pace according to their needs, flip the views of the parts, and disassemble/assemble these parts through the VR training. These good man–system interactions allow the subjects to form a deeper impression of the parts and clearly understand the repair process, which consequently leads to fewer errors and shorter operation time in the performance of the actual maintenance tasks. Another reason could be the difference in MWL between translating from text in the case of TM versus a more direct analogy of the visual matching from the VR to the actual task. Hence, training by VR requires fewer MWL than that of training by TM and contributes to better training performance. In terms of task complexity, the operation time is significantly longer for complex tasks than for simple tasks regardless of the type of training method. These results are reasonable and can be attributed to the higher level of difficulty in parts recognition, as well as the more complicated operation procedures in a complex system. The findings are also consistent with previous literature, which indicated that task complexity influenced the subject's MWL and task performance (Gilbert et al., 2010).

### 4.2 | Effects of training method and task complexity on physiological indices

One purpose of this research is to assess the effects of training method and task complexity on the subject's MWL based on physical indices.

The experimental results indicated that the training method and task complexity significantly affected GSR and AHR of physiological indices. In the training method, the values of GSR and AHR of training by TM are the highest among three methods, followed by MF and VR. No significant difference is observed between the two latter methods. Hence, the MWL of the subjects trained by TM is higher than those of the other two training methods. With regard to task complexity, the values of GSR and AHR are higher in the complex tasks than those in simple tasks. This result is similar to previous studies that obtained higher values of AHR and GSR, which indicated heavier MWL (Knaepen et al., 2015; P. Shi, Hu, & Yu, 2015; Y. Shi, Ruiz, Taib, Choi, & Chen, 2007; Veltman & Gaillard, 1996; Wilson, 2002) and is consistent with the study of Gilbert and his colleagues (2010), which indicated that a complex task would lead to a higher MWL. In the values of HRV-associated indices (LF%, HF%, LF/HF ratio), LF% is significantly affected by the training method and task complexity, and LF/HF ratio is also significantly affected by the training method rather than by task complexity. Both values of LF% and LF/HF ratio of the training by TM are the highest among three methods, followed by MF and VR; however, the result showed no significant difference between MF and VR. The higher LF% and LF/HF ratio mean more severe MWL (Cinaz, Arnrich, Marca, & Troster, 2013; Hwang et al., 2008; Lean & Shan, 2012). Hence, the MWL of training by TM is higher than that of training by MF and VR. The results of the physiological measurements (GSR, AHR, LF%, and LF/HF ratio) indicated a clear difference between the reading manuals (TM) and the use of computer (MF, VR) but did not distinguish between VR and MF. The reasons of not having any difference between the VR and the MF could be attributed to the task complexity or training time, which was neither complicated nor long enough. If the task complexity could be increased or training time could be prolonged, then their difference could be significant. This finding of no difference between the VR and the MF, which is similar to the result of operation errors, implies that simple tasks do not require a complicated training method. This result is consistent with C. H. Lin's study (2014), indicating that with simple systems and low complex operations, the 2D training system was sufficient without using the 3D training system, given the slight change in operation performance and visual fatigue. Saposnik et al. (2016) also pointed out that the VR was no better than a simple recreational activity for stroke recovery. However, the effects of training method and task complexity on HF% are not significant, but a trend that TM has the lowest value of HF% of the three training methods is observed. The lower HF% index indicates higher MWL (Hjortskov et al., 2004; Lean & Shan, 2012).

### 4.3 | NASA-TLX subjective workload assessment

The overall and subscale scores of NASA-TLX subjective workload assessment on the training method are not significant. The result indicated that training method had limited effects on mental, physical, temporal demands, performance, effort, and frustration aspects, as well as the overall workload of the subjects. This finding is not similar to physical indices, such as AHR, GSR, LF%, and LF/HF ratio. One reason could be that the physical MWL measurements can provide "real time" evaluation, thus allowing experimenters to retrieve more accurate signals, whereas subject measurement cannot (Tran et al., 2007). Another reason could be the training contents pertaining to basic inspection, dismantling, replacement, and assembly of three training methods, which were all presented by visual interface and these operations were similar. Therefore, the subject's MWL is also similar. The other reason could be that NASA-TLX subjective evaluation provided less accurate estimation because of individual biases and small number of samples (Lean & Shan, 2012). Although none of them was statistically significant, training by VR had the lowest scores in mental demand, performance, and effort subscales. This result may mean that training by VR requires less mental and perceptual activities, such as thinking, deciding, remembering, and searching, less mental effort to accomplish the goals, and feeling more satisfied with their performance. On the one hand, training by VR has the highest scores in physical demand and frustration subscales. This result may also mean that the subjects required more controlling activities, such as moving a mouse or a keyboard to select unclear items or to flip the objects in a different view angle to understand more details of the task. Moreover, the subjects were prone to feel discouraged and stressed when they clicked on some objects but could not open them. On the other hand, training by TM has the highest scores in mental demand, temporal demand, and performance subscales. This outcome is reasonable given that reading TM needed more mental work to comprehend the text, more time to read the text, and finally led to becoming less satisfied with their performances. Finally, the overall scores show that the subjects considered training by TM subjectively to cause the heaviest workload, whereas training by VR caused the least workload.

## 5 | CONCLUSION

This study used operation performance, physiological indices, and subjective questionnaire to evaluate the effects of training method and task complexity on training performance and MWL. The results of the performance measurement show that VR system is considered the best training method compared to TM and MF, particularly in the case of complex tasks, and the performance of simple tasks are better than those of the complex tasks. The results of physiological measurements (GSR, AHR, LF%, and LF/HF ratio) show a difference between reading TM and using computer (MF and VR), whereas the latter has a lower MWL, and the MWL of complex tasks is higher than that of simple tasks in GSR, AHR, and LF% indices. Although no significance is found in NASA-TLX subject questionnaire and HF%, a trend that training by VR has lower MWL than MF and TM is observed. The findings of VR advantages should be limited to the nonimmersive VR system and tasks provided in our experiment. Three categories of VR system, that is, nonimmersive, half-immersive, and immersive systems with a variety of side effects and limitations were presented early.

## REFERENCES

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

Cates, C. U., Lönn, L., & Gallagher, A. G. (2016). Prospective, randomised and blinded comparison of proficiency-based progression full-physics virtual reality simulator training versus invasive vascular experience for learning carotid artery angiography by very experienced operators. *BMJ Simulation and Technology Enhanced Learning*, *2*(1), 1–5.

Cegrra, J., & Chevalier, A. (2008). The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. *Behavior Research Methods*, *40*(4), 988–1000.

Chaffin, D. B. (2009). Some requirements and fundamental issues in digital human modeling. In V. G. Duffy (Ed.), *Handbook of digital human modeling* (pp. 1–10). New York, NY: Taylor & Francis.

Charlton, S. G. (2002). Measurement of cognitive states in test and evaluation. In S. G. Charlton & T. G. O'Brien (Eds.), *Handbook of human factors testing and evaluation* (pp. 97–126). London, England: Lawrence Erlbaum.

Chung, K. H., Shewchuk, J. P., & Williges, R. C. (2002). An analysis framework for applying virtual environment technology to manufacturing tasks. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *12*(4), 335–348.

Cinaz, B., Arnrich, B., Marca, R. L. A., & Troster, G. (2013). Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and Ubiquitous Computing*, *17*(2), 229–239.

Cobb, S. V. G., Nichols, S., Ramsey, A., & Wilson, J. R. (1999). Virtual reality-induced symptoms and effects (VRISE). *Presence*, *8*, 169–186.

Dadashi, N., Stedmon, A., & Pridmore, T. (2013). Semi-automated CCTV surveillance: The effects of system confidence, system accuracy and task complexity on operator vigilance, reliance and workload. *Applied Ergonomics*, *44*(2), 730–738.

Dahlstrom, N., & Nahlinder, S. (2006). A comparison of two recorders for obtaining in-flight heart rate data. *Applied Psychophysiology and Biofeedback*, *33*(3), 273–279.

Do, P. T., Moreland, J. R., Delgado, C., Wilson, K., Wang, X., Zhou, C., & Ice, P. (2013). Effects of 3D virtual simulators in the Introductory Wind Energy Course: A tool for teaching engineering concepts. *Comprehensive Psychology*, *2*, 4–7.

Duarte, E., Rebelo, F., & Wogalter, M. S. (2010). Virtual reality and its potential for evaluating warning compliance. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *20*(6), 526–537.

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., & Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied Ergonomics*, *52*, 95–103.

Farmer, E., & Brownson, A. (2003). Review of workload measurement, analysis and interpretation methods. *European Organisation for the Safety of Air Navigation*, *33*, 1–33.

Fuhua, L., Duffy, V. G., & Su, C. J. (2002). Developing virtual environments for industrial training. *Information Sciences*, *140*, 153–170.

Gilbert, T. C., Leung, G. Y., & 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.

Goldstein, D. S., Bentho, O. M., Park, Y., & Sharabi, Y. (2011). Low-frequency power of heart rate variability is not a measure of cardiac sympathetic tone but may be a measure of modulation of cardiac autonomic outflows by baroreflexes, *Experimental Physiology*, *96*(12), 1255–1261.

Grantcharov, T. P., Kristiansen, V. B., Bendix, J., Bardram, L., Rosenberg, J., & Funch-Jensen, P. (2004). Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *British Journal of Surgery*, *91*(2), 146–150.

Hart, S. (2006). NASA-task load index (Nasa-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, The Netherlands: North Holland.

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Sogaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, *92*, 84–89.

Hue, P., Delannay, B., & Berland, C. (1997). Virtual reality training simulator for long time flight. In R. J. Seidel & P. R. Chantelier (Eds.), *Virtual reality, training's future?* (pp. 69–76). New York, NY: Plenum.

Hwang, S. L., Yau, Y. J., Lin, Y. T., Chen, J. H., Huang, T. H., Yenn, T. C. & Hsu, C. C. (2008). Predicting work performance in nuclear power plants. *Safety Science*, *46*(7), 1115–1124.

Johnson, A., & Widyanti, A. (2011). Cultural influences on the measurement of subjective mental workload. *Ergonomics*, *54*(6), 509–518.

Knaepen, K., Marusic, U., Crea, S., Rodriguez-Guerruro, C. D., Vitiello, N., Pattyn, N., ... Meeusen, R. (2015). Psychophysiological response to cognitive workload during symmetrical, asymmetrical and dual-task walking. *Human Movement Science*, *40*, 248–263.

LaViola, J. J., Jr. (2000). A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, *32*(1), 47–56.

Lean, Y., & Shan, F. (2012). Brief review on physiological and biochemical evaluations of human mental workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *22*(3), 177–187.

Lee, C. H., Chou, C., & Sun, T. L. (2015). Evaluating presence for customer experience in a virtual environment: Using a nuclear power plant as an example. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *25*(4), 484–499.

Lehmann, K. S., Ritz, J. P., Maass, H., Cakmak, H. K., Kuehnapfel, U. G., Germer, C. T., ... Buhr, H. J. (2005). A prospective randomized study to test the transfer of basic psychomotor skills from virtual reality to physical reality in a comparable training setting. *Annals of Surgery*, *241*(3), 442–449.

Lin, C. H. (2014). *The effect of different devices in presenting VR to the assessment of human performance and visual fatigue* (master's thesis). Retrieved from National Digital Library of Theses and Dissertations in Taiwan. http://handle.ncl.edu.tw/11296/ndltd/99367879564801255848

Lin, F., Ye, L., Duffy, V. G., & Su, C. J. (2002). Developing virtual environments for industrial training. *Information Sciences*, *140*(1), 153–170.

Lin, J. W., Duh, H. B. L., Parker, D. E., Abi-Rached, H., & Furness, T. A. (2002). Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment. In *Virtual Reality, 2002*. Proceedings IEEE (pp. 164–171). IEEE. http://handle.ncl.edu.tw/11296/ndltd/99367879564801255848

Mujber, T. S., Szecsi, T., & Hashmi, M. S. (2004). Virtual reality applications in manufacturing process simulation. *Journal of Materials Processing Technology*, *155*, 1834–1838.

Nathanael, D., Mosialos, S., Vosniakos, G. C., & Tsagkas, V. (2016). Development and evaluation of a virtual reality training system based on cognitive task analysis: The case of CNC tool length offsetting. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *26*(1), 52–67.

Ogden, G. D., Levine, M., & Eisner E. J. (1979). Measurement of workload by secondary tasks. *Human Factors*, *21*, 529–548.

Osberg, K. M. (1995). Virtual reality and education: Where imagination and experience meet. *VR in the School*, *1*(2), 1–3.

Ragan, E. D., Bowman, D. A., Kopper, R., Stinson, C., Scerbo, S., & McMahan, R. P. (2015). Effects of field of view and visual complexity on virtual reality training effectiveness for a visual scanning task. *IEEE Transactions on Visualization and Computer Graphics*, *21*(7), 794–807.

Reyes del Paso, G. A., Langewitz, W., Mulder, L. J., Roon, A., & Duschek, S. (2013). The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: A review with emphasis on a reanalysis of previous studies. *Psychophysiology*, *50*(5), 477–487.

Robertson, G. G., Card, S. K., & Mackinlay, J. D. (1993). Three views of virtual reality: Nonimmersive virtual reality. *Computer*, *26*(2), 81.

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*(1), 61–86.

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, *35*(11), 991–1009.

Saposnik, G., Cohen, L. G., Mamdani, M., Pooyania, S., Ploughman, M., Cheung, D., ... Nilanont, Y. (2016). Efficacy and safety of non-immersive virtual reality exercising in stroke rehabilitation (EVREST): A randomised, multicentre, single-blind, controlled trial. *The Lancet Neurology*, *15*(10), 1019–1027.

Shi, P., Hu, S., & Yu, H. (2015). Influence of computer work under time pressure on cardiac activity. *Computers in Biology and Medicine*, *58*, 40–45.

Shi, Y., Ruiz, N. R., Taib, E., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2651–2656). New York, NY: ACM.

Sztajzel, J. (2004). Heart rate variability: A noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Medical Weekly*, *134*, 514–522.

Tran, T. Q., Boring, R. L., Dudenhoeffer, D. D., Hallbert, B. P., Keller, M. D., & Anderson, T. M. (2007, August). Advantages and disadvantages of physiological assessment for next generation control room design. In *2007 IEEE 8th Human Factors and Power Plants and HPRCT 13th Annual Meeting* (pp. 259–263). IEEE.

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

Vilar, E., Rebelo, F., & Noriega, P. (2014). Indoor human wayfinding performance using vertical and horizontal signage in virtual reality. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *24*(6), 601–615.

Wang, L. M., Duffy, V. G., & Du, Y. (2007). A composite measure for the evaluation of mental workload. In V. G. Duffy (Ed.), *Digital human modeling* (pp. 460–466). Berlin, German: Springer.

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) (pp. 965–969). SAGE Publications.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology*, *12*(1), 3–18.

Wu, X., Mu, G., Yang, Z. J., & Gu, C. (2012). Design and implementation of interactive virtual maintenance training system for tank gun. In *Computer Science and Electronics Engineering (ICCSEE)*, *3*, 383–387.

**How to cite this article:** Chao C-J, Wu S-Y, Yau Y-J, Feng W-Y, Tseng F-Y. Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance. *Hum Factors Man*. 2017;27:187–196. https://doi.org/10.1002/hfm.20702