When simulated environments make the difference: the effectiveness of different types of training of car service procedures.

Simone Borsci

Diagnostic Evidence Cooperative of London, Department of Surgery and Cancer, Imperial College London, St Mary's Hospital, South Wharf Road, W2 1NY London, United Kingdom. e-mail: simone.borsci@gmail.com. Corresponding author.

Glyn Lawson

Human Factors Research Group, Faculty of Engineering, The University of Nottingham, Nottingham NG7 2RD, United Kingdom. e-mail: Glyn.Lawson@nottingham.ac.uk

Davide Salanitri

Human Factors Research Group, Faculty of Engineering, The University of Nottingham, Nottingham NG7 2RD, United Kingdom. e-mail: Davide.Salanitri@nottingham.ac.uk

Bhavna Jha

Jaguar Land Rover, Abbey Road, Whitley, Coventry, CV3 4LF, United Kingdom. e-mail: bjha1@jaguarlandrover.com

Accepted version: 23 April 2016 (online first) for Virtual Reality of Springer

Abstract

An empirical analysis was performed to compare the effectiveness of different approaches to training a set of procedural skills to a sample of novice trainees. Sixty-five participants were randomly assigned to one of the following three training groups: i) learning-by-doing in a 3D desktop virtual environment, ii) learning-by-observing a video (show-and-tell) explanation of the procedures, and iii) trial-and-error. In each group participants were trained on two car service procedures. Participants were recalled to perform a procedure either two or four weeks after the training. The results showed that: i) participants trained through the virtual approach of learning-by-doing performed both procedures significantly better (i.e., $p < .05$ in terms of errors and time) than people of non-virtual groups. ii) The virtual training group, after a period of non-use, were more effective than non-virtual training (i.e., $p < .05$) in their ability to recover their skills. iii) After a (simulated) long period from the training – i.e., up to 12 weeks – people who experienced 3D environments consistently performed better than people who received other kinds of training. The results also suggested that independently from the training group, trainees' visuospatial abilities were a predictor of performance, at least for the complex service procedure, Adj $R^2 = .460$, and that post-training performances of people trained through virtual learning-by-doing are not affected by learning styles. Finally, a strong relationship ($p<..001$ $R^2 = .441$) was identified between usability and trust in the use of the virtual training tool – i.e., the more the system was perceived as usable, the more it was perceived as trustable to acquire the competences.

Keywords

Automotive; Car service maintenance; Training effectiveness; Training evaluation; Virtual Reality

1. Introduction

Virtual environments (VE) are simulated and 3D systems in which users are immersed and able to perceive and interact with a three-dimensional world (Bowman et al. 2002; Milgram and Kishino 1994). VEs are used daily in several fields, e.g., manufacturing, surgery, education, and military, as part of programmes to train employees, specialists, and managers in a variety of skills and skilled tasks (e.g., prototype and assembly, drive, fight, fly, surgery procedures etc. see: Belardinelli et al. 2008; Jayaram et al. 2007; Seth et al. 2011). VE tools are usually applied as an extension of classic learning activities for the enhancement of professionals' procedural skills, intended as the kinds of training programmes to: i) covert, through practice, trainees' knowledge into procedural form, and ii) to help trainees acquire the ability to perform all the actions of a procedure correctly and proficiently – i.e., in the correct order and within a time limit (Anderson 1982; Fitts 1967).

In the automotive field the application of VEs as training systems is relatively new. Recent research projects show that automakers today are struggling to find new technological systems to support learning activities for car service operators (see for instance, European projects SKILLS - http://www.skills-ip.eu/ and VISTRA - http://www.vistra-project.eu). Automakers are fully aware that training people in procedural skills for car service maintenance is very complex for at least two main reasons. First, cars are usually customized to suit customers' needs. Therefore, the same model of a car may vary substantially in terms of components and internal electronic design. For instance, a luxury car model can have up to 10^{24} possible configurations – e.g., different engine, chassis, electronic configurations (Parry et al. 2011). Consequently, an operator who has to change the same fault component on two cars, with different customizations, could face different procedures and different disassembly and reassembly steps. Second, operators trained on the service procedures for a newly released model of car, are typically required to service this model several months after receiving their training. In fact, in the automotive industry there is usually a large temporal delay between the training of operators' competences and the real execution of the service procedures on a car. For instance, training of service operators could be rolled out several weeks ahead of the launch of a new vehicle, with limited opportunity for operators to see a real car. However, service teams may only see the vehicle several months later – namely when a customer experiences some issue with their new car.

Automakers, to support car service operators' acquisition of competences, have invested in two main types of learning activities performed during 'in presence' training: i) learning-by-observing (i.e., observational learning, see: Bandura 1992), in which trainees acquire the theory and the main steps of the procedures by looking at experts performing the procedure, and ii) learning-by-doing, (i.e., also known as experiential learning, see: Kolb 1984) in which trainees apply the new skills during simulations of the procedure.

Nevertheless, the classic centralized training programmes which combine observational and experiential learning are expensive, and usually require that operators leave their work premises and current jobs for the period of the training. Moreover, due to the costs, 'in presence' training is usually given to a limited number of operators, who after the training share the new know-how with their colleagues in service centres.

Automakers, by aiming to minimize the costs associated with 'in presence' activities, and to increase the possibility that operators could access new know-how (such as an update on a service procedure), have invested in online systems in which operators can receive long-distance theoretical training, and download paper manual, apps and video training (Alippi et al. 2003; Anastassova and Burkhardt 2009). So far, limited effort has been focussed on tools to support learning-by-doing activities (Anastassova et al. 2005; Stork et al. 2012; Xu and Gao 2011). Rather, automotive industries have focused their efforts on the development of tools to support service operators' observational learning – i.e., instruction papers, video manuals, online courses, and gamified tools. Today, to fill the lack of long-distance experiential learning training, automakers are interested in VE tools as systems which could allow people to actually exercise and simulate their competences at their work premises.

1.1. Effectiveness of training with VE systems in different fields

As Borsci, Lawson and Broome (2014) underline, industrial companies are used to developing their own VE solutions which serve specific training programmes for specific roles (e.g., managers, suppliers, trainers, trainees etc. Anastassova et al. 2005). Therefore, operators are trained by each company to perform variable procedures – e.g., assembly of, or maintenance on, different products (Michalos et al. 2010). In this context, the effectiveness criteria of the training programmes are determined by the companies, thus producing a high level of uncertainty about the effective application of VE systems. Industrial research on the effectiveness of VE is far behind compared to the analyses performed in specialized fields of VE applications, for example healthcare, surgery, military and aviation (Li et al. 2003; Yuviler-Gavish et al. 2013; Yuviler-Gavish et al. 2011a; Yuviler-Gavish et al. 2011b). Four main gaps between industrial and specialized application of VEs for training operators can be highlighted in the literature (Borsci et al. 2014). First, in specialist training the effectiveness of VE systems have been measured through standardized tools, by testing several indices of performances to check the proficiency level reached by trainees. On the other hand, effectiveness of VE training in industrial fields were mainly assessed only by measuring (before, during and after the training) two indices of operators' performances: i) the number of errors, and ii) the time taken by operators to perform the procedure (Tang et al. 2003; Webel et al. 2013; Yuviler-Gavish et al. 2013; Yuviler-Gavish et al. 2011a). Second, researchers in industry have not measured consistently trainees' achievement of proficiency after the training (Ahlberg et al. 2007). This index is particular important in the automotive sector, because to satisfy customers of a service centre, operators have to be able to repair a car correctly and in certain range of time. Third, in the industrial field no studies previously analysed and compared the ability of trainees to recover the acquired procedural skills after a period of time from training, despite the well-established phenomena of skill decay (Arthur Jr et al. 1998; Arthur et al. 1997) which is considered an important measure of training effectiveness (Stefanidis et al. 2005; Wayne et al. 2006). Finally, in line with the recent review paper of Borsci, Lawson and Brome (2014), researchers in industry rarely gather data, through standardized tools, on those factors which could affect the both acquisition of the skills and the interaction with VE tools, such as:

i) Trainees' previous abilities – e.g., their visuospatial abilities (Ahlberg et al. 2007; Garg et al. 2002; Parsons et al. 2004), or mental rotation abilities (Peters et al. 1995).

ii) Trainees' attitudes, styles and expectations – such as learning styles (Ai-Lim Lee et al. 2010), expectations and trust in the use of technologies (Flavián et al. 2006; Mcknight et al. 2011).

iii) Complexity of the training tasks, and workload perceived by trainees (Stefanidis et al. 2007; Stone et al. 2011).

iv) Trainees' satisfaction in the use and usability of the VE system using standardised scales (assessment is often performed using qualitative methods).

v) Trainees' cybersickness during the interaction with a VE system (Sharples et al. 2008). As Borsci and colleagues (2014) outlined, cybersickness is not always measured, or reported, in empirical analysis of industrial applications of VE systems for training (e.g., see: Li et al. 2003; Webel et al. 2013; Yuviler-Gavish et al. 2013; Yuviler-Gavish et al. 2011b)

1.2. Aim of the study

The present study compares the effectiveness of training car service operations using three different approaches: learning-by-doing activities on a VE system, video training and free learning. The comparative analysis was performed using a LEGO® model to simulate car service procedures.

Our experimental design was developed to explore the following two hypotheses. The first hypothesis is based on previous evidence, from surgery and specialized fields, which showed that VE training is more effective than video and trial-and-error approaches (e.g., see: Hamilton et al. 2002; Nagendran et al. 2013).

Hypothesis 1: The type of training received by people significantly affects their abilities to perform the procedure immediately after the training. In particular, after a learning-by-doing training through a VE people perform a procedure in less time and with fewer errors than trainees who received training by video, or who received no training at all.

Associated to this hypothesis, we proposed another one, as follows:

Hypothesis 1-1. The type of training received by people significantly affects their abilities to recover the acquired skills after a period of time from the knowledge acquisition. In particular, people who acquired the know-how through a learning-by-doing approach in a VE system recover their abilities quicker and better (less time and errors to perform proficiently) than people trained through an observational or a trial-and-error approach.

In fact, while only a few comparative studies on VE systems assessed trainees' abilities to recover their skills after the period of time (Carlson et al. 2015; Hall et al. 1998), these studies suggest that people trained through VE systems can recover the know-how to proficiently perform a procedure, quicker and better than people exposed to other types of training.

The second hypothesis is based on convincing indications that an unusable system leads to a low level of trust in use, thus compromising the attitude of users toward a product – i.e., people do not want to use an unreliable system (Christine Roy et al. 2001; Lippert and Michael Swiercz 2005; Mcknight et al. 2011; Sasse 2005). Currently, no studies explore the relationship between usability and trust in the field of VE. To explore this relationship we used two validated instruments: the trust in technology measures (TTM, Mcknight et al. 2011), and the System Usability Scale (SUS, for a review, see: Borsci et al. 2009; Borsci et al. 2013; Brooke 1996; Lewis 2014; Lewis and Sauro 2009). Our hypothesis is that:

Hypothesis 2. The more a VE system is perceived as usable, the more people trust the use of that system. In particular, we expect to observe a correlation between SUS and TTM overall scales, and their sub-factors of these scales.

Finally, for each group of participants we also explore the effects on people's performances immediately and after a period of time of the following seven factors: people characteristics and previous experience, cybersickness, complexity of the procedure, types of instruction selected/used to perform, visuospatial abilities, learning style, cognitive workload.

2. Method

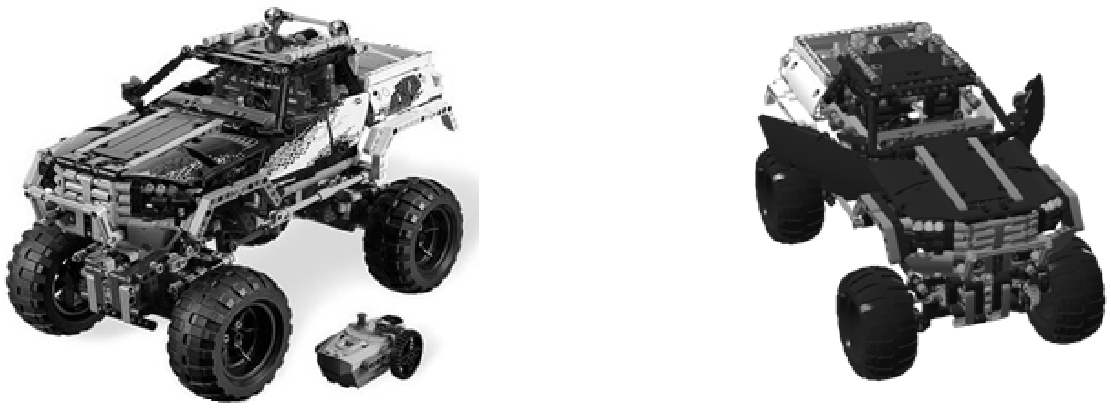
We conducted the study in two phases. In phase 1, participants were randomly assigned to different kinds of training (i.e., trial-and-error, training in a VE, or training through video observation) to learn how to perform two car service procedures with different levels of complexity, i.e., with different number of steps and different proficiency time limits. The performances of trainees after the training were used to compare the training effectiveness, and to observe the effect on people's performances of both procedure complexity and people's previous abilities. In phase 2, all participants were asked to perform, after a period of time from the training (two or four weeks), only the second procedure that they learned in phase 1. The outcomes of phase 2 were used to observe and compare the effectiveness of training (with or without VE systems) with regards to people's ability to recover their acquired skills.

2.1. Participants

Data were collected between May and July 2014 at The University of Nottingham. A sample of 65 participants (32 Male, Age: 30.07, SD: 7.99) was recruited by paper and online advertisements, and were paid in £20 shopping vouchers to compensate their inconvenience. All the participants were volunteers for both phase 1 and 2. Participants were recruited from staff and students of high schools and universities in the Nottinghamshire area. Demographics data are shown in Appendix 1.

2.2. Tools and Apparatus

To simulate the complexity of the car systems experienced by technicians in daily service procedures a real and a virtual model of LEGO ® Technic 4X4 Crawler (n. 9398) composed of 1327 bricks was used for the two experiments – see Figure 1.



a. Real model of car

b. Virtual model of car

Figure 1. LEGO ® Technic 4X4 Crawler car. Section a, shows a picture of the real model of car – weight of 1.420g, high 21cm, long 39cm and wide 21cm. Section b, shows a picture of the virtual reproduction of the model.

We used the LEGO ® Technic model as a metaphor of a real procedure for two main reasons: First, in the scientific literature, several studies have used LEGO® models to simulate and compare skills acquisition (Ottosson 2002; Yuviler-Gavish et al. 2011b); Second, LEGO® offered a free 3D tool in which people can operate on virtual models – i.e., LEGO® Digital Designer version 4.3 (LDD), which we used as tool for training each participant on two service procedures . The objective of the first procedure was to swap the central engine of the car (i.e. the battery) with a new one (Appendix 2). The second procedure required to participants to swap the left front damper of the car with a new one (Appendix 2).

The first procedure was composed of 36 physical steps – e.g., open car, remove/insert components – and 13 manipulative steps – e. g, remember how to rotate/move the car. The second procedure was composed by 64 physical and 25 manipulative steps. In this paper, we will refer to this complex service task as the 'target procedure' (see Figure 2).

<table>
<tr><th></th><th colspan="2">Phase 1</th><th colspan="2">Phase 2</th></tr>
<tr><td>AIMS</td><td colspan="2">Training effectiveness, and factors which affect people performances</td><td colspan="2">people's ability to recover their acquired skills</td></tr>
<tr><td rowspan="5">PROCEDURE</td><td colspan="2">Pre-experimental survey (see section 2.5)</td><td rowspan="2">Two weeks from phase 1</td><td rowspan="2">Four weeks from phase 1</td></tr>
<tr><td>Procedure 1 Training</td><td>Target procedure Training</td></tr>
<tr><td colspan="2">Manipulation of a real 4X4 Crawler car</td><td colspan="2">Manipulation of a real 4X4 Crawler car</td></tr>
<tr><td>• Trial
• Test</td><td>• Trial
• Test</td><td rowspan="2">Test target procedure</td><td rowspan="2">Test target Procedure</td></tr>
<tr><td colspan="2">Post-experimental measures (see section 3.3)</td></tr>
<tr><td>PARTICIPANTS</td><td colspan="4">65 volunteers</td></tr>
<tr><td>PROFICIENCY LIMITS</td><td colspan="4">Time limits of procedures, list of errors (section 2.3)</td></tr>
<tr><td>INDICATORS OF PERFORMANCES</td><td colspan="4">Performance indexes (see section 2.4)</td></tr>
</table>

Figure 2. Study paradigm diagram. The design of the research consists of two different phases (phase 1 and 2). In phase 1, participants were asked to fill the pre-experimental tools and then they were trained on two service procedures. After the training, participants were asked to perform the procedures on a real model of 4X4 Crawler (trial and test), and to fill-out the post experimental

measures. In phase 2 participants were asked to perform only the target procedure after 2 or 4 weeks from the training.

The experimental scenarios were presented to participants on a flipchart board. Moreover participants could, on the basis of their preferences, use as a support tool one of the following two kits of instructions: i) a paper manual with pictures of the real Lego® Crawler4x4 model during each step and explanation of how to perform correctly each procedure; or ii) a video manual containing a video of the real Lego® Crawler4x4model during the procedure and verbal explanation of each step. A desktop computer with Microsoft © Windows 7 Enterprise, processor Intel © core i7, 3.70 GHz, 8Gb of Ram and a dedicated graphics and sound system was used by participants to interact with the LDD environment and to explore the video manual. An iPad mini with 16Gb of RAM was used by trainers to show time and countdown to the users during the experiment, and to collect notes. Moreover, a Sony HDD DCR-SR57 video camera was used to record the participants.

2.3. Proficiency time and list of performance errors

To set up the proficiency time for first and target procedure, three voluntary engineers with long, self-declared expertise in the use of LEGO ® Technic were trained face-to-face on the two procedures through a show-and-tell explanation by using the real Lego® model of the 4X4 Crawler. Experts were asked to perform the procedures as fast as they could without errors, for at least five times – only attempts without errors were counted as valid procedures. The average time of experts' performances was used to define a proficiency time limit for each procedure – 102 seconds for the first procedure, and 121 seconds for the target procedure.

In previous literature, some researchers of training in the industrial field (Webel et al. 2013; Yuviler-Gavish et al. 2013; Yuviler-Gavish et al. 2011a) have discriminated between two kind of errors performed by trainees after the training: 'solved/solvable' errors, i.e. errors performed but solved by operators, and 'unsolved/unsolvable', i.e. errors performed and left unsolved by trainees. While this approach is a useful way to discriminate among different kinds of errors we designed our experiments to avoid 'unsolvable errors'. In fact, any time participants performed an error they were informed by trainers and forced to solve the issue alone, or through the use of instruction. We used this approach because car operators, in real service procedures, are allowed to use instruction during the service to avoid errors. In line with that, in this study, an error was identified during participants' performances when one of the following issues was noticed by trainers:

- Participants needed to use the instructions to perform correctly e.g., they did not remember how to proceed, they performed wrongly one or more steps of the procedure without correcting these mistakes, they forget to perform one or more steps or delayed their actions for more than 10 seconds.
- Participants did not follow correctly the procedure but they identified on their own the solution to the problem, e.g., participants performed wrongly or forgot one or more steps of the procedure, but they immediately corrected the mistake; participants missed one or more components by losing time to recover the bricks etc.

2.4. Performances indices

We assessed and compared participants' performances in the two procedures using the following seven indices:

i) *Trial errors*: Number of errors performed during the trial after the training;

ii) *Trial time*: Time (sec) spent to perform the trial after the training;

iii) *Attempts*: Number of attempts needed to proficiently perform the procedure. This index was estimated as the sum of attempts performed by a participant until they perform correctly the procedure within the time limit;

iv) *Errors*: Overall number of errors. This index was obtained as the sum all the errors the participant made in all the attempts;

v) *Time*: Overall time to perform proficiently, obtained as the sum of time (sec) spent by participants in all the attempts;

vi) *Av. Time*: Average time (sec) spent to perform proficiently. This index was obtained as a ratio between index *v* and *iii*;

vii) *Av. Errors*: Average number of errors – this index was obtained as a ratio between index *iv* and *iii*.

2.5. Pre and post experimental survey

Before the training, each participant filled both consent and demographic forms (14 questions – see Appendix 3a). We used these data to collect users' characteristics (such as age, gender, education etc. see Q1-Q5, Appendix 3a), their previous experience with LEGO® Technic (Q6), 3D virtual environments (Q7) videogames (Q8), LDD tool (Q9), and their previous experience of cybersickness using interactive tools (Q7a-Q9a).

Each participant was also required to complete the following two standardized tools:

- A test of mental rotation abilities (i.e., version A of redrawn mental rotation, MRA, see: Peters et al. 1995).
- A learning style inventory (KLSI, version 3.1., composed of 12 items, see: Kolb and Kolb 2005) to determine their learning style.

Finally, only participants who received VE training were requested to fill before and after each use of the LDD tool a Simulator Sickness Questionnaire (SSQ, see: Kennedy et al. 1993). SSQ is composed of a list of 16 symptoms asking the participants to mark, in scales from 1 – None to 4 – Severe, if any of the symptoms apply to them.

After the test of the target procedure (performed on a real model of 4X4 Crawler), participants of the three groups were asked to fill-out the NASA-TLX (Hart and Staveland 1988) to assess their workload in executing this procedure. Moreover, only participants of VT group were asked to rate their satisfaction and trust in the use of LDD. As stated before, to explore and to observe the relationship between trust and usability, we used two validated instruments: the TTM developed by McKnight et al. (2011) – composed of seven factors: Reliability, Functionality, Helpfulness, Situational Normality, Structural Assurance , Faith in General Technology and Trusting Stance – and the SUS, developed by Brooke (1996) – composed of two factors: Usability and Learnability.

3. Experiment design

In phase 1, immediately after the pre-experimental phase each of the 65 participants were randomly assigned to one of the following training groups:

- Free learning group (FL): Participants of this group did not receive any training on the two procedures. Trainers only explained through written scenarios the aim of each procedure, but also showed the participants how the instruction kits were organized. Participants of FL group were invited to acquire the procedures through a process of trial-and-error.
- Classic training group (CT): The aim of each procedure was presented to participants as a car service scenario (see Appendix 2). Trainees then received an explanation on how to use the instruction kits. After that, participants were invited to watch a movie in which an expert performed and verbally explained each step of the procedures on the real Lego® model of the 4X4 Crawler. People could ask only once to rewind the video after each step of the procedure. Participants of this group acquired the procedure through a process of learning-by-observing (Bandura 1992).
- Virtual tool training group (VT): Participants received training (5 minutes) on the principal functions of LDD system, but also performed five minutes of free exercises of assembly and disassembly on a virtual car composed of 523 bricks available on the LDD gallery (http://ldd.lego.com/en-us/gallery/archive). The aim of each procedure was presented to trainees as a service scenario (see Appendix 2), and they received instructions about the use of the instruction kits. After that, participants were asked to perform each procedure on the virtual model of the 4X4 Crawler in the LDD system following a set of virtual instructions. These participants only interacted with a virtual version of the car without seeing the real Lego® model of the 4X4 crawler. However, different to video training, people of the VT group performed a learning-by-doing activity (Kolb 1984).

Each group was developed, in line with the indications of automotive experts at Jaguar Land Rover, to represent different simulated training programmes of car service. The FL group represents the scenario in which operators have to learn a new procedure alone by relying only on instructions. The

CT group represents the case of long distance training with a low cost support – i.e., a video training. The VR group represents the case of long distance training with high cost support – i.e. a dedicated virtual environment. We did not use a condition in which people are trained by observing or manipulating a real model of the Lego® 4X4 Crawler because, as explained earlier, this condition represents the centralized 'in presence' training programmes that companies would like to reduce by using technologies for long distance training. To support the trial-and-error procedure, but also the participants of CT and VT, trainers showed the two instruction kits (i.e., paper and video), and explained to participants that they could use only one of the two instruction kits as a support tool when they had problems in performing a procedure. Participants were then invited to select on the basis of their preferences one of the two kits (see Appendix 3b).

After training on each of the two procedures, independently from the group they belong to, participants had access to a real model of the Lego® 4X4 Crawler and were asked to perform one trial and one test, as follows:

1. Trial: Each participant was requested to perform correctly, without a time limit, the procedure they have just acquired (first or target). Participants were also informed that their performances were timed, but they were instructed, if they did not remember how to perform, to ask to stop the time to look at the instructions kit. Nevertheless, participants may ask for the instructions kit only after each 30 seconds. Moreover, participants were informed that at any time during the performance, whether they performed a mistake without solving it, the trainer can stop the time and forced them to use the instructions. Trainers were instructed to remind participants (when they showed some uncertainties) that they could stop the time and ask for the instructions.
2. Test: After the trial, participants were requested to perform against the time limit associated with each procedure. Participants were informed that they must perform within the proficiency time limit, otherwise after they had finished the procedure they have to restart from the beginning. For each procedure, participants had three attempts to perform below the threshold. Moreover, participants were also informed that they may ask for the instructions kit after each 30 seconds, but that the timer could not be stopped. Finally, any time that a

participant performed out of the procedure, trainers were instructed to read this statement aloud: “You have to perform a correct procedure. At the moment you have made a mistake, try to solve it, otherwise you cannot go further with the procedure. You can also check the instruction if you want, but I cannot stop the time”.

After the performance of the two procedures, participants were asked to perform the post experimental survey and participants were randomly allocated to one of two sub-groups for the second experimental phase. The sub-group determined the period of time – two or four weeks – from phase 1 after which participants were invited to perform a re-test of the target procedure – i.e., with the same tools and experimental procedure. Participants in phase 2 had a maximum of 3 attempts to perform the target procedure, below the proficiency time limit, or below their best performance during the test.

4. Results

4.1. Data analysis

Both in phase 1 and 2, t-tests, and analysis of variance (ANOVA) were used to test the effect of the participants’ abilities and reactions to the training approaches (pre and post experimental measures) on people’s performances during and after training. Moreover, bootstrap linear regression was applied to simulate the outcomes of the three training groups with a resampled cohort of 1000 people (Akins et al. 2005; Good 2000). In this study, we applied the bootstrap to check the goodness of the outcomes of the regressions performed on the original sample. Finally, a forecasting analysis was performed through an exponential smoothing with Holt’s linear trend method to simulate the expected trend of people performances (in terms of errors and time) after several weeks from the training i.e., up to 12 weeks.

Descriptive statistics, bootstrapped linear regression analysis, ANOVA and t-test analysis were used to analyse and compare participants’ characteristics and performances of the two procedures – i.e., indices from *i* to *viii (Section 2.4)*.

The analysis was performed by IBM®SPSS 22.

Results are discussed below with reference to the study aims and hypotheses.

4.2. Hypothesis 1: Immediately after the training, the virtual group performs better than other groups.

ANOVA analysis showed a significant difference among performances of the participants of the three training groups (Table 1).

Table 1. ANOVA analysis of the significant differences (*p*) between the indexes of performances in the two procedures performed by VT, CT and FL groups, and descriptive statistics mean (M) and standard deviation (SD).

Indexes of performance		**Difference in performances**		**FL**		**CT**		**VT**	
First procedure		F (5, 65)	*p*	Mean	*SD.*	Mean	*SD.*	Mean	*SD.*
Trial	*iii) Trial errors*	16.4	.001	7.18	*3.47*	5.2	*3.1*	2.4	*1.5*
	iv) Trial time	--	--	224.7	*81.4*	210.2	*72.7*	195	*48*
Test	*v) Attempts*	3.15	.05	1.4	*.74*	1.3	*.57*	1.05	*.21*
	vi) Errors	11.35	.001	4.2	*3.6*	3.2	*1.8*	.82	*1.1*
	vii) Time	--	--	255.8	*631.6*	119.6	*58.6*	79.1	*23.8*
	viii) Av. time	--	--	204.2	*637.9*	77.1	*20.58*	75.5	*15.5*
	ix) Av. errors	14.05	.001	2.7	*1.5*	2.4	*1.3*	.82	*1.2*
Target procedure		F (5, 65)	*p*	Mean	*SD.*	Mean	*SD.*	Mean	*SD.*
Trial	*i) Trial errors*	52.53	.001	12	*2.69*	9.8	*3.7*	3.3	*2.2*
	ii) Trial time	4.42	.016	373.3	*163.7*	382.2	*103.3*	280.8	*94.5*
Test	*iii) Attempts*	12.11	.001	3.3	*.99*	3.3	*.956*	2.04	*.92*
	iv) Errors	13.16	.001	14.3	*9.5*	10.8	*6.71*	3.6	*3.6*
	v) Time	13.12	.001	412.8	*151.5*	408.2	*139.7*	228.1	*114*
	vi) Av. time	--	--	74.9	*7.2*	79.8	*15*	85.2	*19*
	vii) Av. errors	10.46	.001	4.1	*2.4*	3.5	*1.5*	1.7	*1.5*

LSD post-hoc analysis showed that participants of VT group demonstrated more effective performance (i.e., less errors, indexes *i* and *iv*) than participants of the FL and CT groups. In particular, in the first procedure, participants of VT group performed significant less errors (index *iv*) than participants of FL (p=.002) and CT group (p=.001). In the target procedure, VT group demonstrated a significant difference (p=.001) with FL and CT, both in terms of errors of performance (indexes *i* and iv) but also in terms of time (index *v*).

4.3. Hypothesis 1.1: After a period of time virtual group will recover the acquired skills better than other groups.

ANOVA analysis showed differences among the performances of the groups after two and four weeks (Table 2).

Table 2. ANOVA analysis of the significant differences (*p*) between the indexes of performances of FL, CT, and VT group during the retest of procedure.

Indexes of performance		Difference in performance	
		F	**p**
Re-test of target procedure	*x) Attempts*		
	xi) Errors	3.306	.011
	xii) Time	2.275	.05
	xiii) Av. time	1.796	.128
	xiv) Av. errors	4.136	.003

In tune with our hypothesis LSD post-hoc analyses showed that:

- After two weeks, people of the VT group performed a lower overall number of errors (iv) than people who do not received structured training (FL, p=.049), or who experienced classic training (CT, p=.04). Concurrently, participants of the VT group performed lower errors per attempts (vii) than (FL, p= .001) and CT (p=.031).
- After 4 weeks, the group trained by the VE system performed better than the other groups in terms of number of errors (iv) than participants of FL (p=.016) and CT (p=.041), and with less errors per attempt (vii) than participants of FL (p=.045) and CT (p=.031). Moreover, the

VT group participants were able to proficiently reacquire their ability by spending less time to perform than other groups, i.e., FL (p=.039) and CT (p=.013).

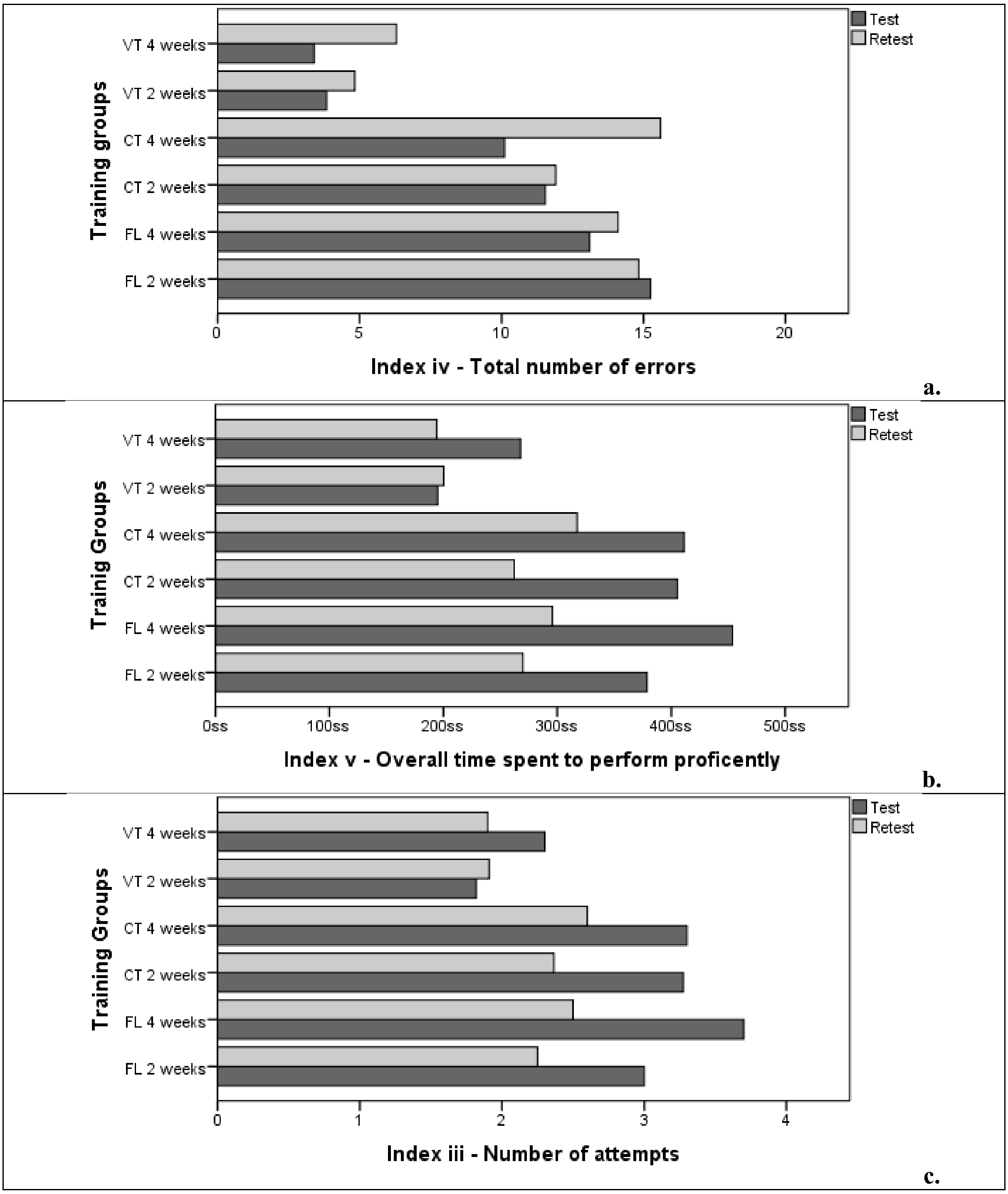


Figure 3. Representation of the differences between the participants' performances in the three groups measured after two and four weeks. Performances were measured as errors (index *iv*, Figure

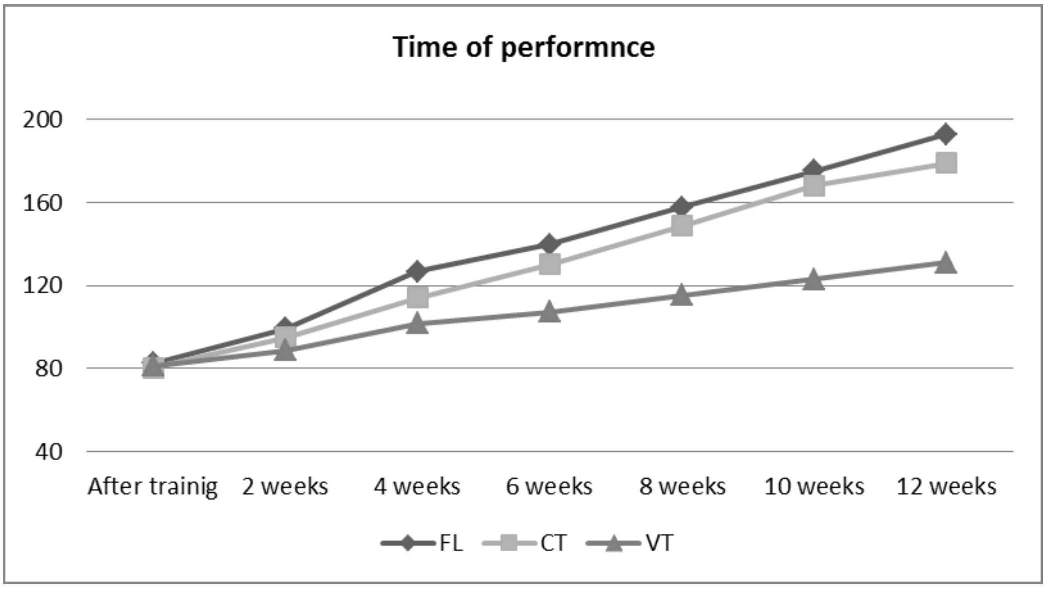
3a), times (index *v*, Figure 3b) and number of attempts (index *iii*, Figure 5c) spent by participants to perform the target procedure.

To simulate the performances of the three groups after a long period from training – i.e., 12 weeks – we applied an exponential smoothing with Holt's linear trend method on two sets of data: the average number of errors, and average time to perform proficiently. To compose our datasets we used the participants' performances gathered immediately after the training (to represent the starting point of a trained operator), and after two and four weeks. This analysis allowed us to simulate the ability of people, who have received different kinds of training, to apply the acquired skills after a long period of non-use.

Figure 4 shows the forecasted values and the graphical representation of the performances of each group over a period of 12 weeks, in terms of time (Figure 4, section a) and errors (Figure 4, section b).

a. Time of performance forecasted values

Indexes of performances	Group	RMSE	MAE	MPEA	M
vi) Av. time	FL	27.51	12.95	11.04	139.09
	CT	7.84	3.67	3.28	130.65
	VT	13.47	6.38	6.54	106.89



b. Errors forecasted values

Indexes of performances	Group	RMSE	MAE	MPEA	Overall Mean
vii) Av. errors	FL	.178	.877	14.81	6.82
	CT	.336	.156	3.224	5.59
	VT	.823	.367	15.81	2.95

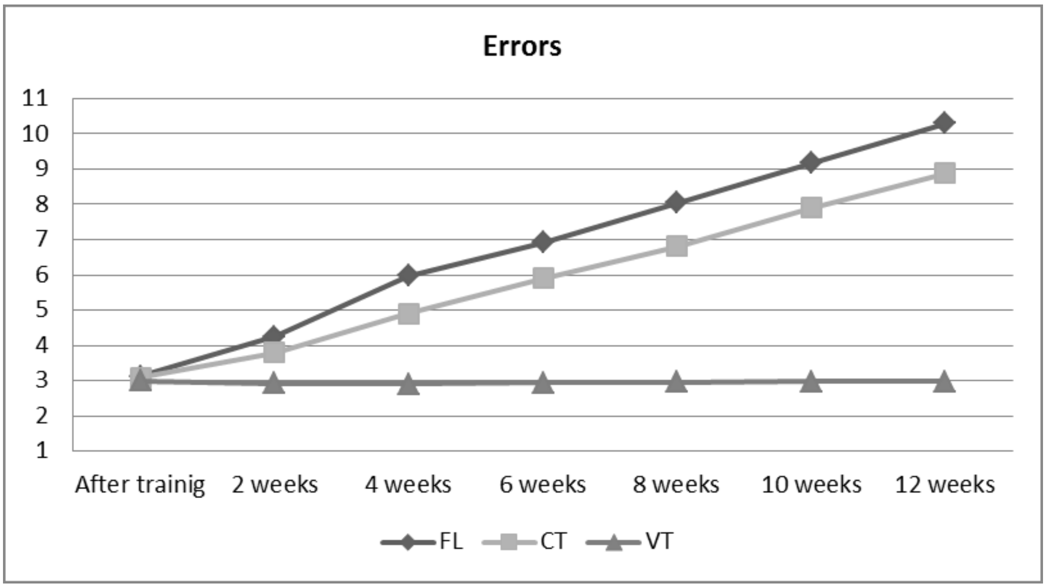


Figure 4. Forecasted distribution of errors and time needed to perform proficiently procedure 2 from immediately after the training up to 12 weeks for the competences acquisition. Root-mean-square

error (RMSE), mean absolute error (MAE), mean absolute percentage error (MPEA) and the overall mean (M) of the distribution were calculated for each training group: FL, CT and VT.

4.4. Hypothesis 2: The more people perceived a virtual system as usable, the more people trust the use of that system

A linear regression analysis showed neither SUS nor TMM scores affected the participants' performances during the trial in terms of errors and time (indices i and *ii*). Nevertheless, as shown in Table 3, participants of VT group, through TTM and SUS scales, judged LDD as a trustable and satisfactory system.

Table 3. Mean and standard deviation of factors and overall score of trust in technology measures (TMM) and system usability scale (SUS).

Scales	**Factors***	**Mean****	**SD**
TTM			
	Reliability	70.18	11.06
	Functionality	75.97	14.28
	Situation normality	81	10.65
	Structural assurance	59.68	19.05
	Faith in technology	74.83	15.91
	Trusting Stace	70.18	13.82
Overall TTM score		**73.07**	**8.96**
	Factors	**Mean**	**SD**
SUS			
	Usability	57.87	9.5
	Learnability	15.79	4.1
Overall SUS score		**73.63**	**10.08**

*We excluded from the list of TTM factors the Helpfulness because participants did not used any help tool during the interaction with LDD.
** We transform for our use the TTM outcomes in percentage

Pearson's correlation coefficient (r) analysis suggested that overall trust in the LDD system (i.e., overall TTM score) was correlated with the overall satisfaction of use measured by SUS (r=.527, p=.012), and strongly correlated with the usability factor of SUS (r=.664, p=.001). Moreover, the usability factor of SUS correlated with two TMM factors: functionality (r=.472, p=.027) and reliability (r=.467, p=.028). A linear regression indicated that the usability factor of SUS is a significant predictor of overall trust of participants (p=.001, R^2=.441) – i.e., the more the system is perceived as a usable interactive tool the more it is considered a trustable training technology.

Moreover, clear tendencies, though not significant, are displayed in the data about the relationship among SUS overall scale and usability factor, and TMM overall score.

4.5. Factors which affect people's performances when trained through different approaches

People characteristics and previous experience: The level of education (Q5) significantly affected the number of attempts and the time to proficiently perform target procedure. Moreover, experience in video games (Q8) and with Lego (Q6) marginally affects participants' performances (see Table 4).

Table 4. Effect of personal characteristics of participants' performances in the two procedures.

Indexes of performance	**Predictors**								
	Education (Q5)			**Lego experience (Q6)**			**Game (Q8)**		
First procedure	***B***	***P***	**Adj R^2**	***B***	***p***	**Adj R^2**	***β***	***p***	**Adj R^2**
iv) Errors	-	-	-	-.450	.001	.190	-.447	.001	.200
Target procedure	***B***	***P***	**Adj R^2**	***B***	***p***	**Adj R^2**	***β***	***p***	**Adj R^2**
i) Trial errors	-	-	**.321**	-	-	.120	-.463	.001	.201
iii) Attempts	-.308	.010		-	-		-	-	
iv) Errors	-	-		-.346	.001		-	-	
v) Time	-.242	.040		-	-		-	-	

Adjusted $R^2 \geq .300$ are bolded

Bootstrapped linear regression analysis confirmed that experience with Lego® (Q6) affected the number of errors during the trial of the first procedure of the FL group (index *i*, R^2=.300), and the time needed to become proficient to perform target procedure of the VT group (index *v*, R^2=.372).

Moreover, education level (Q5) affected only participants of the CT group in terms of number of attempts to perform proficiently the target procedure (index *iii*, R^2=.308). On the other hand no effects were noted for experience with games (Q8).

Cybersickness: The t-test analysis of the pre- and post-cybersickness questionnaires showed that participants who interacted with the LDD tool did not experience any motion-sickness problems during the interaction with the system.

Types of instruction selected/used to perform: 24.6% of the cohort selected a paper manual, instead of a video one, to receive support during the procedure. In particular, 45% of people in the FL group chose the paper manual instead of the video one, while only 22.7% of the VT group and 9.6% of the CT group preferred the paper over video manual. An ANOVA performed among people's performances (indices) and the type of instructions suggested that there were no significant differences between the performances of participants who decided to use a paper or a video manual. However, a clear tendency was displayed in the data in which people who selected video manual performed both the procedures in less time and with fewer errors, than people who selected a paper manual.

Visuospatial abilities: The average' score of the MRA cohort was equal to 11.15 (SD: 3.74). An ANOVA analysis underlined that there were no significant differences among FL, CT and VT groups in terms of mental rotation abilities. Table 5 showed that MRA scores were significant predictors of participants' errors during the trial, number of attempts needed, and time to be proficient to perform the target procedure immediately after the training.

Table 5. Effect of mental rotation abilities on participants' performances.

Indexes of performances	**Predictors** **MRA**		
First procedure	***B***	***p***	**Adj R^2**
i) Trial errors	-.349	.001	.264
iii) Attempts	-.297	.012	
Target procedure	***B***	***p***	**Adj R^2**
i) Trial errors	-.605	.001	**.460**
iii) Attempts	-.531	.001	
v) Time	-.794	.001	

Adjusted $R^2 \geq .300$ are bolded

The bootstrapped regression analysis within the three training groups showed that MRA significantly affected people's performances, independently from the training, as follows:

- index *iii*, the number of attempts need to perform proficiently the first procedure of the FL group – i.e., R^2=.311.
- index *iv*, the number of errors performed by participant during test of the target procedure of the FL (R^2=.331) and VT (R^2=.301) groups;
- index *v*, the time to proficiently perform target procedure of the FL (R^2=.324) and CT (R^2=.300) groups.

Learning style: The analysis of the KLSI revealed that the majority of participants involved in this study have a converging learning style (36.9%), followed by assimilators (27.7%), divergers (18.5%) and accommodators (17.9%). The ANOVA analysis showed that there were no significant differences among the three training groups in terms of people learning styles. Regression analysis showed that learning style did not affect people's performances measured immediately after the training, or after two and four weeks.

Cognitive Workload: The mean of NASA-TLX overall scale for all the participants was equal to 68.81 (Min.:41.33; Max.: 94.67; SD: 12.75). An ANOVA was performed among the indices of performances and participants overall score of NASA-TLX. Results showed that the overall time (index *v*) to perform target procedure was significantly associated with the cognitive workload perceived by participants (F (3,65)=4.551, p=.007) – i.e., the higher the level of perceived workload, the higher the time spent by users to perform the procedure immediately after the training. Regression analysis confirmed that, independently from the training group, the time for people to perform proficiently the target procedure significantly (p=.001) affected the workload perceived by participants (R^2=.304). This effect was also identified within the participants of VT (R^2=.302), CT (R^2=.377) groups, and FL group (R^2=.225).

Finally to explore which factor affected the skills recovery in terms of time and errors (after two and four weeks), we performed a linear regression analysis. The analysis revealed that mental rotation ability and perceived workload (i.e., MRA and NASA-TLX scores) affect the recovery of acquired skills after a period of time (see Table 6), as follows:

- After two weeks, the perceived cognitive workload had no effect on people performances, while MRA scores of people slightly predicts abilities to recover their acquired skills.
- After 4 weeks, both MRA and NASA-TLX scores strongly affected people’s performances in terms of errors – i.e., the higher the MRA scores, the lower are the number of errors (index *iv*), and the lower are the NASA-TLX scores, the lower are the errors to perform target procedure.

Table 6. Effect of mental rotation abilities (MRA), and cognitive workload (NASA-TLX) perceived by people on the retest of target procedure after two and four weeks.

Index of performances	**2 weeks**			**4 weeks**				
	MRA			**MRA**		**NASA-TLX**		
	B	***p***	**Adj R2**	***β***	***p***	***β***	***P***	**Adj R2**
iii) Attempts	*-.533*	*.001*	262	*-364*	*.028*	.417	.013	**.352**
v) Time	*-.490*	*.003*	217					

5. Discussion

This research supports the idea that, to train car service operators, virtual systems in which operators may acquire a knowledge trough a learning-by-doing experience of training, are better approaches that video instructions or trial-and-error. Table 7 summarizes the testing outcomes for the hypotheses.

Table 7. Summary of Test Outcomes for Hypotheses of the Empirical Analysis

Hypotheses	**Result**	**Meaning**
Hypothesis 1	Supported	Immediately after the training, people who acquired car service procedures through a learning-by-doing simulation in a VE perform better (i.e., less time and errors) than people trained by video or trial-and-errors.
Hypothesis 1-1	Partially Supported	The type of training received by people significantly affects their abilities to recover the acquired skills after a period of time. People trained through a VE system performed after two and four weeks better than people of CT and FL groups. However after two weeks from the training, the difference among the groups is only in terms of errors of performance, while after four weeks the people trained through VE can recover the skills to perform a complex car service procedure by spending less time and performing less errors than people trained by video and trial-and-errors.
Hypotheses 2	Supported	Overall trust, functionality and reliability (TTM) of the system are strongly related to the usability. Moreover, usability factor of SUS is a significant predictor of the overall trust measured through TTM.

The results of Hypothesis 1 showed, in line with previous research (Nagendran et al. 2013), that immediately after training in a VE system, people perform procedures in less time and with less errors than people trained through observational or trail-and-error approaches. Moreover, as shown by the results of Hypothesis 1-1, the greater the time from the training, the greater the differences in skill recovery among the groups (Carlson et al. 2015). However after two weeks, the VE group performed better than other groups only in terms of errors, while after four weeks a clear difference among the groups emerged in terms of time and errors of performance.

It should be noted that the results may be specific to the configurations used in this study. As mentioned in Section 3: Experiment Design, the groups were developed with automotive industry experts to reflect current approaches to training in automotive service tasks in industry. Thus, while the benefits seen in the VE group are likely to be realised in other similar training tasks (i.e. procedural tasks) this may not be the case when compared to groups that differ with regards to the levels of instruction, observation and practise as used in this study. This is an interesting area of future research: to explore the generalisability of the findings and the level of benefits which can be attributed solely to the training media, rather the training approach as a whole.

The results of Hypothesis 2 indicate that people who perceived a VE system as usable are more likely to trust the system as a reliable and functional tool of training. To develop effective VE training systems automakers have to invest not only in the design of the functions of the VE systems, but also in usability assessment of these products. In fact the usability of VE systems is a key factor to improve the overall experience of the users, and to positively affect trainees' perception of the system. This is because, the more a training system is perceived as usable, the more trainees believe that the system is reliable – i.e., a consistent tool of training (Mcknight et al. 2011) – and functional – i.e., a tool able to deliver, properly. a new set of new competences (Mcknight et al. 2011).

6. Conclusion

Our outcomes highlight that training delivered through a VE tool was more effective in terms of skills acquisition (performance after the training) and skills recovery (after a short and long period from the training) compared to non-virtual approaches to training. By simulating car service procedures performed on a LEGO ® Technic model of car, the present study allowed us to compare the effectiveness of different kinds of virtual and non-virtual training approaches. The overall results support the idea that for manufacturers to invest in VE systems, simulating a learning-by-doing experience of a car service procedure may substantially enhance operators' performances. Moreover, the outcomes showed that VE systems are more effective than other kinds of training to increase operators' abilities to recover, after a period of non-use, the acquired competences. Nevertheless, the initial investment to develop a usable and trustable VE training system is still higher (in terms of costs and time) than, for instance, the costs to create video trainings or to make online courses. Therefore, auto-makers, on the basis of their needs, have to define a balance between virtual and non-virtual approaches of training for service operators, by aiming to achieve the best trade-off between costs and benefits in the use of VE systems.

Further studies are needed to replicate our comparative analysis by testing people's training and performances on a real car. Therefore, our main future aim is to develop a multiplatform virtual tool –

i.e., one that may be used on different kind of hardware, such as: CAVE, Oculus Rift, interactive tables and desktop PCs – for training operators on real car service procedures. This kind of system will help us to extend our current outcomes, and also to compare the effectiveness of training delivered through different kind of devices.

Acknowledgements

This paper was completed as part of Live Augmented Reality Training Environments (LARTE) – 101509 project. The authors would like to acknowledge the Technology Strategy Board for funding the work.

References

Ahlberg G et al. (2007) Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies The American Journal of Surgery 193:797-804 doi:http://dx.doi.org/10.1016/j.amjsurg.2006.06.050

Ai-Lim Lee E, Wong KW, Fung CC (2010) How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach Computers & Education 55:1424-1442 doi:http://dx.doi.org/10.1016/j.compedu.2010.06.006

Akins R, Tolson H, Cole B (2005) Stability of response characteristics of a Delphi panel: application of bootstrap data expansion BMC Medical Research Methodology 5:37

Alippi C, de Russis C, Piuri V (2003) A neural-network based control solution to air-fuel ratio control for automotive fuel-injection systems Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 33:259-268 doi:10.1109/TSMCC.2003.814035

Anastassova M, Burkhardt J-M (2009) Automotive technicians' training as a community-of-practice: Implications for the design of an augmented reality teaching aid Applied Ergonomics 40:713-721 doi:http://dx.doi.org/10.1016/j.apergo.2008.06.008

Anastassova M, Burkhardt J-M, Mégard C, Ehanno P (2005) Results from a user-centred critical incidents study for guiding future implementation of augmented reality in automotive maintenance International Journal of Industrial Ergonomics 35:67-77 doi:http://dx.doi.org/10.1016/j.ergon.2004.08.005

Anderson JR (1982) Acquisition of cognitive skill Psychological Review 89:369-406

Arthur Jr W, Bennett Jr W, Stanush PL, McNelly TL (1998) Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis Human Performance 11:57-101 doi:10.1207/s15327043hup1101_3

Arthur JW, Day EA, Bennett JW (1997) Dyadic Versus Individual Training Protocols: Loss and Reacquisition of a Complex Skill Journal of Applied Psychology 82:783-791

Bandura A (1992) Observational learning. In: Squire L (ed) Encyclopedia of learning and memory. Macmillan, New York, USA,

Belardinelli C, Blümel E, Müller G, Schenk M (2008) Making the virtual more real: research at the Fraunhofer IFF Virtual Development and Training Centre Cogn Process 9:217-224 doi:10.1007/s10339-008-0216-0

Borsci S, Federici S, Lauriola M (2009) On the Dimensionality of the System Usability Scale (SUS): A Test of Alternative Measurement Models Cogn Process 10:193–197 doi:10.1007/s10339-009-0268-9

Borsci S, Kurosu M, Federici S, Mele ML (2013) Computer Systems Experiences of Users with and without Disabilities: An Evaluation Guide for Professionals. CRC Press, Boca Raton, FL. doi:10.1201/b15619-1

Borsci S, Lawson G, Broome S (2014) Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance Computers in Industry:Accepted, in press

Bowman DA, Gabbard JL, Hix D (2002) A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods Presence: Teleoperators and Virtual Environments 11:404-424 doi:10.1162/105474602760204309

Brooke J (1996) SUS: A "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL (eds) Usability evaluation in industry. Taylor & Francis, London, UK, pp 189–194

Carlson P, Peters A, Gilbert SB, Vance JM, Luse A (2015) Virtual Training: Learning Transfer of Assembly Tasks Visualization and Computer Graphics, IEEE Transactions on 21:770-782 doi:10.1109/TVCG.2015.2393871

Christine Roy M, Dewit O, Aubert BA (2001) The impact of interface usability on trust in Web retailers Internet Research 11:388-398 doi:doi:10.1108/10662240110410165

Fitts PM, & Posner, M. I. . (1967) Human performance. Brooks/Cole, Monterey, CA

Flavián C, Guinalíu M, Gurrea R (2006) The role played by perceived usability, satisfaction and consumer trust on website loyalty Information & Management 43:1-14 doi:http://dx.doi.org/10.1016/j.im.2005.01.002

Garg AX, Norman GR, Eva KW, Spero L, Sharan S (2002) Is There Any Real Virtue of Virtual Reality?: The Minor Role of Multiple Orientations in Learning Anatomy from Computers Academic Medicine 77:S97-S99

Good P (2000) Permutation tests: a practical guide for resampling methods for testing hypotheses. Springer-Verlag, New York, NY

Hall CR, Stiles RJ, Horwitz CD Virtual reality for training: evaluating knowledge retention. In: Virtual Reality Annual International Symposium, 1998. Proceedings., IEEE 1998, 18-18 1998 1998. pp 184-189. doi:10.1109/VRAIS.1998.658488

Hamilton EC et al. (2002) Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills Surg Endosc 16:406-411 doi:10.1007/s00464-001-8149-z

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Peter AH, Najmedin M (eds) Advances in Psychology, vol Volume 52. North-Holland, pp 139-183. doi:http://dx.doi.org/10.1016/S0166-4115(08)62386-9

Jayaram S, Jayaram U, Kim YJ, DeChenne C, Lyons KW, Palmer C, Mitsui T (2007) Industry case studies in the use of immersive virtual assembly Virtual Real 11:217-228 doi:10.1007/s10055-007-0070-x

Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness The International Journal of Aviation Psychology 3:203-220 doi:10.1207/s15327108ijap0303_3

Kolb AY, Kolb DA (2005) The Kolb Learning Style Inventory—Version 3.1 2005 Technical Specifications. Hay Group Holdings, Inc., Boston

Kolb DA (1984) Experiential Learning as the Science of Learning and Development. Prentice Hall, Englewood Cliffs, NJ

Lewis JR (2014) Usability: Lessons Learned … and Yet to Be Learned International Journal of Human-Computer Interaction 30:663–684 doi:10.1080/10447318.2014.930311

Lewis JR, Sauro J (2009) The Factor Structure of the System Usability Scale. In: Kurosu M (ed) Human Centered Design, vol 5619. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp 94–103. doi:10.1007/978-3-642-02806-9_12

Li JR, Khoo LP, Tor SB (2003) Desktop virtual reality for maintenance training: an object oriented prototype system (V-REALISM) Computers in Industry 52:109-125 doi:http://dx.doi.org/10.1016/S0166-3615(03)00103-9

Lippert SK, Michael Swiercz P (2005) Human resource information systems (HRIS) and technology trust Journal of Information Science 31:340-353 doi:10.1177/0165551505055399

Mcknight DH, Carter M, Thatcher JB, Clay PF (2011) Trust in a specific technology: An investigation of its components and measures ACM Trans Manage Inf Syst 2:1-25 doi:10.1145/1985347.1985353

Michalos G, Makris S, Papakostas N, Mourtzis D, Chryssolouris G (2010) Automotive assembly technologies review: challenges and outlook for a flexible and adaptive approach CIRP Journal of Manufacturing Science and Technology 2:81-91 doi:http://dx.doi.org/10.1016/j.cirpj.2009.12.001

Milgram P, Kishino F (1994) A Taxonomy of Mixed Reality Visual Displays IEICE Transactions on Information Systems E77-D:1321-1329

Nagendran M, Gurusamy KS, Aggarwal R, Loizidou M, Davidson BR (2013) Virtual reality training for surgical trainees in laparoscopic surgery Cochrane Database of Systematic Reviews doi:10.1002/14651858.CD006575.pub3

Ottosson S (2002) Virtual reality in the product development process Journal of Engineering Design 13:159-172 doi:10.1080/09544820210129823

Parry G, Newnes L, Huang X (2011) Goods, Products and Services. In: Macintyre M, Parry G, Angelis J (eds) Service Design and Delivery. Service Science: Research and Innovations in the Service Economy. Springer, New York, US, pp 19-29. doi:10.1007/978-1-4419-8321-3_2

Parsons TD, Larson P, Kratz K, Thiebaux M, Bluestein B, Buckwalter JG, Rizzo AA (2004) Sex differences in mental rotation and spatial rotation in a virtual environment Neuropsychologia 42:555-562 doi:http://dx.doi.org/10.1016/j.neuropsychologia.2003.08.014

Peters M, Laeng B, Latham K, Jackson M, Zaiyouna R, Richardson C (1995) A Redrawn Vandenberg and Kuse Mental Rotations Test - Different Versions and Factors That Affect Performance Brain and Cognition 28:39-58 doi:http://dx.doi.org/10.1006/brcg.1995.1032

Sasse M (2005) Usability and Trust in Information Systems. In: Mansell RaC, B (ed) Trust and Crime in Information Societies. Edward Elgar, Cheltenham, UK, pp 319 - 348

Seth A, Vance J, Oliver J (2011) Virtual reality for assembly methods prototyping: a review Virtual Reality 15:5-20 doi:10.1007/s10055-009-0153-y

Sharples S, Cobb S, Moody A, Wilson JR (2008) Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems Displays 29:58-69 doi:http://dx.doi.org/10.1016/j.displa.2007.09.005

Stefanidis D et al. (2007) Construct and face validity and task workload for laparoscopic camera navigation: virtual reality versus videotrainer systems at the SAGES Learning Center Surgical Endoscopy 21:1158-1164 doi:10.1007/s00464-006-9112-9

Stefanidis D, Korndorffer Jr JR, Sierra R, Touchard C, Dunne JB, Scott DJ (2005) Skill retention following proficiency-based laparoscopic simulator training Surgery 138:165-170 doi:http://dx.doi.org/10.1016/j.surg.2005.06.002

Stone RT, Watts KP, Zhong P, Wei C-S (2011) Physical and Cognitive Effects of Virtual Reality Integrated Training Human Factors: The Journal of the Human Factors and Ergonomics Society 53:558-572 doi:10.1177/0018720811413389

Stork A, Sevilmis N, Weber D, Gorecky D, Stahl C, Loskyll M, Michel F Enabling virtual assembly training in and beyond the automotive industry. In: Virtual Systems and Multimedia (VSMM), 2012 18th International Conference on, 2-5 Sept. 2012 2012. pp 347-352. doi:10.1109/VSMM.2012.6365944

Tang A, Owen C, Biocca F, Mou W Comparative effectiveness of augmented reality in object assembly. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA, 2003. ACM, 642626, pp 73-80. doi:10.1145/642611.642626

Wayne DB, Siddall VJ, Butter J, Fudala MJ, Wade LD, Feinglass J, McGaghie WC (2006) A Longitudinal Study of Internal Medicine Residents' Retention of Advanced Cardiac Life Support Skills Academic Medicine 81:S9-S12

Webel S, Bockholt U, Engelke T, Gavish N, Olbrich M, Preusche C (2013) An augmented reality training platform for assembly and maintenance skills Robot Auton Syst 61:398-403 doi:10.1016/j.robot.2012.09.013

Xu W, Gao F Research of virtual reality technology in automotive engine assembly teaching. In: Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International, 20-22 Aug. 2011 2011. pp 167-169. doi:10.1109/itaic.2011.6030177

Yuviler-Gavish N, Gutiérrez T, Webel S, Rodríguez J, Peveri M, Bockholt U, Tecchia F (2013) Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks Interactive Learning Environments:1-21 doi:10.1080/10494820.2013.815221

Yuviler-Gavish N, Gutierrez T, Webel S, Rodriguez J, Tecchia F (2011a) Design Guidelines for the Development of Virtual Reality and Augmented Reality Training Systems for Maintenance and Assembly Tasks BIO Web of Conferences 1:29-21, 29-24 doi:10.1051/bioconf/20110100029

Yuviler-Gavish N, Yechiam E, Kallai A (2011b) Learning in multimodal training: Visual guidance can be both appealing and disadvantageous in spatial tasks International Journal of Human-Computer Studies 69:113-122 doi:http://dx.doi.org/10.1016/j.ijhcs.2010.11.005