Behaviour & Information Technology, 2014
Vol. 33, No. 1, 16–35, http://dx.doi.org/10.1080/0144929X.2012.681067

Taylor & Francis
Taylor & Francis Group

The impact of self-efficacy and perceived system efficacy on effectiveness of virtual training systems

Dawei Jia*, Asim Bhatti and Saeid Nahavandi

Centre for Intelligent Systems Research (CISR), Deakin University, Geelong, Australia

(*Received 13 July 2009; final version received 27 March 2012*)

This study developed and tested a research model which examined the impact of user perceptions of self-efficacy (SE) and virtual environment (VE) efficacy on the effectiveness of VE training systems. The model distinguishes between the perceptions of one's own capability to perform trained tasks effectively and the perceptions of system performance, regarding the established parameters from literature. Specifically, the model posits that user perceptions will have positive effects on task performance and memory. Seventy-six adults participated in a VE in a controlled experiment, designed to empirically test the model. Each participant performed a series of object assembly tasks. The task involved selecting, rotating, releasing, inserting and manipulating 3D objects. Initially, the results of factor analysis demonstrated dimensionality of two user perception measures and produced a set of empirical validated factors underlining the VE efficacy. The results of regression analysis revealed that SE had a significant positive effect on perceived VE efficacy. No significant effects were found of perceptions on performance and memory. Furthermore, the study provided insights into the relationships between the perception measures and performance measures for assessing the efficacy of VE training systems. The study also addressed how well users learn, perform, adapt to and perceive the VE training, which provides valuable insight into the system efficacy. Research and practical implications are presented at the end of the paper.

Keywords: human–virtual environment interaction; learning outcomes; training evaluation; performance; perceptions; memory

1. Introduction

It is apparent at all managerial levels and in all functional areas that there is currently an increased attention and interest given to the utilisation of advanced computer technologies. One of such technologies, virtual environment (VE), is at the core of many training, education and entertainment platforms due to its potential to enhance the human ability to learn abstract concepts and complex procedural tasks. A VE refers to a computer-generated, 3D spatial environment based on the real-world or abstract objects and data. It possesses the features of 3D immersion, multisensory cures, frames of reference (FORs, are spatial metaphors which can enhance the meaningfulness of data and provide qualitative insights) and employs an advanced human-computer interface (including advanced displays) as well as modes of interaction to engage multiple human sensorial channels (e.g. visual, auditory/hearing, and haptic/touch) during an interaction experience (Bowman *et al*. 2004). During recent years VE has become a promising tool for training and education (Brough *et al*. 2007). Despite its adaptation for training and fast-paced technological advancements, the ways in which to evaluate efficacy of such technology are unclear.

Much research attentions have been given to usability evaluation of 2D computer technologies; however, not many well-established methods for 3D VEs evaluation are reported. It has been argued that the traditional usability evaluation methods need to be altered to better suit evaluating 3D VEs (Stanney 2002, Stanney *et al*. 2003, Bowman *et al*. 2004). Recently, many studies (Nichols *et al*. 2000, Lin 2007, Theng *et al*. 2007) have paid specific attention to VE evaluation, aimed at achieving a better understanding of user experience and usability and to consider to what extent users are able to use the technology effectively, efficiently and with satisfaction. Most of the research activity has focused on identifying factors which relate to the interaction experience, such as immersion, presence, engagement and control; or to the human factor issues, such as simulator sickness, cognitive load or after effect. Many studies have only relied on task performance measures for quantifying the effectiveness of VEs. Nevertheless, it is unlikely that a single evaluation criterion is adequate to capture the complexity of VE efficacy. Specifically, when designed for training, such as the training of complex procedural tasks, it is important that a more complete set of

*Corresponding author. Email: dwj@deakin.edu.au

© 2012 Taylor & Francis

factors contributing to a learner/trainee's ability to learn the skills, are taken into account in evaluation.

1.1. Learning outcomes

Assessment of the effectiveness of a training programme requires systematic collection of data to clearly demonstrate learning outcomes after training. Kirkpatrick (1959) suggests there are four types of outcomes which can account for the effectiveness of training. These outcomes are reaction, learning, behaviour, and results.

- Reactions – trainee feelings, attitudes, and opinions about training
- Learning – the skills and knowledge acquired in training
- Behaviour – the transfer of learned skill and knowledge to the workplace
- Results – the impact of training on the organisation in terms of cost reduction, quality improvement, increases in quantity of work, and reduced absenteeism

Similarly, Kraiger *et al.* (1993) claim that learning is a multidimensional construct and may be evident from changes in one's cognitive, skill-based and affective capacities in a training programme. Therefore, training evaluation needs to take a construct-oriented approach that measures learning in terms of cognitive, affective and skill-based outcomes and involves a systematic collection of data that measures multidimensional learning outcomes to quantify the success of training programmes. They further suggest that cognitive learning outcomes focus on the dynamic processes of knowledge acquisition, organisation and application. Affective learning outcomes pertain to the learners' motivation, attitudes, feelings and opinions about training. Skill-based learning outcomes refer to the development of technical or motor skills. Even though these outcomes appear to have different terminologies, i.e. 'affective' and 'skill-based' and 'cognitive' learning outcomes as defined by Kraiger *et al.* (1993), and 'reactions', 'learning' defined by Kirkpatrick (1959), they share the same conceptual meaning. Specifically, both 'affective' and 'reactions' are subjective perceptions or responses from the trainee/learner of the training system. 'Learning' and 'skill-based' outcome refer to the learners' performance of trained skills, and the knowledge acquired ('cognitive learning) in training. This study utilised theories of cognitive, skill-based and affective learning outcomes for training evaluation (Kraiger *et al.* 1993).

1.2. Training evaluation for quantification of VE efficacy

Kraiger *et al.* (1993) claim that self-report measures, such as self-rating on self-efficacy (SE) or perceived system performance capability, are the most appropriate methods for the evaluation of the affective dimension. Targeted behavioural observation, hands-on testing, and structured situational interviews are useful methods in evaluating skill development in training (Ostroff 1991, Marcolin *et al.* 2000, Sue-Chan and Ong 2002). In terms of measuring skill-based learning, speed of performance, error rates, fluidity of performance that reflects composition proceduralisation use are all effective (Yoo and Bruns 2005, Tavakoli *et al.* 2006). Moreover, secondary task performance could also be used to assess a trainee's cognitive learning that requires automatic processing and tuning of learnt skills (Willingham 1998, Yang *et al.* 2008). In terms of evaluation of cognitive learning, methods such as self-reporting, recognition and recall tests could be used (Kraiger *et al.* 1993, Lin 2004, 2007).

Based on these training evaluation methods, and past studies of VE evaluation (North *et al.* 2001, Bowman *et al.* 2002, Lin 2004, Popovici and Marhan 2008), affective learning outcomes were examined with respect to SE and perceive VE efficacy. Skill-based outcomes were examined in terms of performance test of trained skills in VEs. Cognitive-based outcomes can be examined through a performance memory test. These measures were grouped into subjective perceptual based measures, i.e. SE and perceived VE efficacy (PVE), and objective performance-based measures, i.e. performance of a training test and performance of a memory test, which were designed to evaluate VE efficacy.

2. Research model and hypotheses

Figure 1 presents the conceptual framework within which the proposed model was formulated. Based on theories of cognitive, skill-based and affective learning outcomes (Kraiger *et al.* 1993), the framework argues that affective learning outcomes (user perceptions of SE and PVE) are theorised to influence cognitive (memory) and skill-based learning outcomes (task performance). Specifically SE may affect memory either directly or indirectly through task performance, and a positive relationship will be found between PVE and task performance. In addition, task performance is theorised to influence memory and mediate the effects of SE and PVE. Finally, the model posits that SE may also affect PVE. The theoretical rationale for the model draws upon training evaluation research from both within and beyond the computer-based simulation and

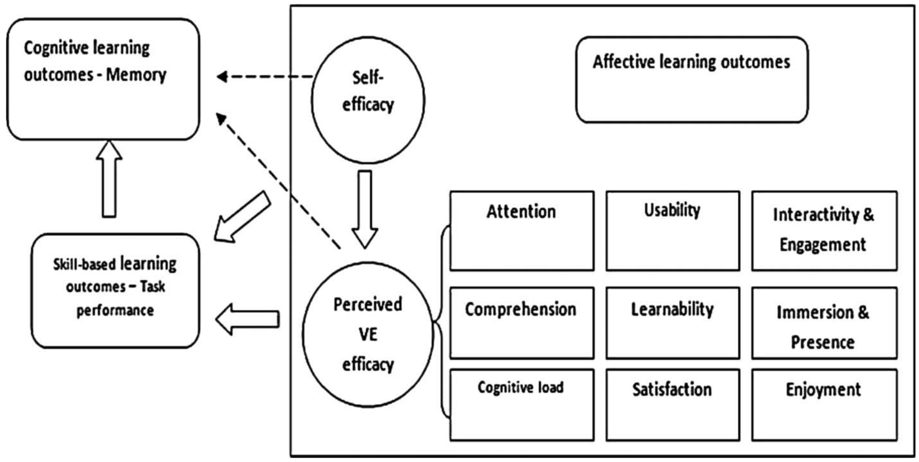

Figure 1. Conceptual framework.

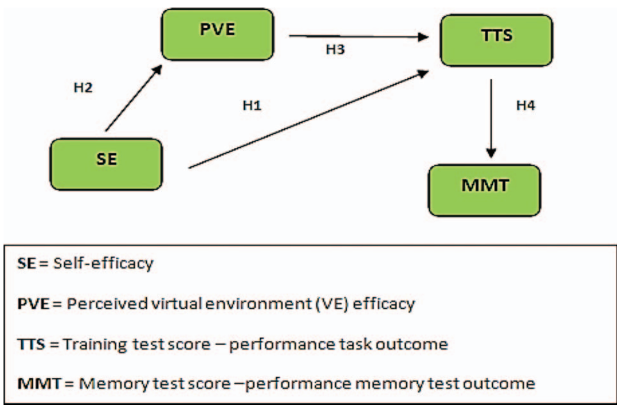

Figure 2. Graph depicting the relationships among perceptions, performance and memory.

Table 1. Hypotheses summary.

Hypothesis	
H1	Self-efficacy will have a positive effect on performance task outcome
H2	Self-efficacy will have a positive effect on perceived VE efficacy
H3	Perceived VE efficacy will have a positive effect on task outcome
H4	Performance task outcome will be positively related to recognition and recall in a performance memory test
H5	No direct relationship will be found between perceived VE efficacy and memory
H6	No direct relationship will be found between SE and memory

training domain. The model is specifically intended to apply within the domain of VE procedural tasks training (e.g. object assembly), and is not designed to be generalised beyond these boundary conditions (e.g. to non-computer training). Construct of the model also infers reference to human–computer interaction, user interface design and system evaluation literature as well as empirical HCI studies. Figure 2 shows a graph depicting the relationships between perceptions, performance and memory that further specifies each element of the proposed model examined in this study and the hypotheses (as Table 1 illustrates) relating them.

2.1. *Self-efficacy*

SE refers to an individual's expectancy in his/her capability to organise and execute the behaviours needed to successfully complete a task (Bandura 1997).

Figure 3. Employed virtual training environment for the proposed evaluation study.

It has been conceptualised both as an antecedent to training and an outcome of training (Yi and Davis 2003). Also SE has been used to predict decision making, cognitive task performance, and mathematical test scores (Wang and Newlin 2002), as well as proving to be beneficial in problem-solving efficiency (Riley *et al.* 2004). Importantly, research has shown that SE shares a moderate relationship with knowledge acquisition and the subsequent task performance (Gist 1997). In addition, post-training SE has proven to be a useful predictor of cognitive learning and long-term skill maintenance (Yi and Davis 2003), as well as subsequent task performance (Johnson and Marakas 2000), and therefore should be included as a post-training measure of learning (Kraiger *et al.* 1993). Likewise, (Yi and Davis 2003) it was found that post-training SE had a significant effect on both immediate and delayed task performance.

Prior research has also shown that SE beliefs and attitudes towards a computer are indicators of performance in computer-mediated learning (North *et al.* 2001). Johnson (2005) also claimed that strong evidence was found regarding the importance of SE and performance relationship in computer skill acquisition. According to Hasan (2008), SE represents the amount of effort and persistence that people exert to perform a task successfully; therefore, it is hypothesised that individuals with higher SE beliefs are expected to expend more effort to understand and learn the new skill. As a result they will demonstrate higher levels of learning performance than those with lower SE beliefs. Based on Hasan's (2008) recent study that investigated the impact for SE on the acquisition of computer skills, a significant and positive effect of SE on far-transfer of learning has been found. Far-transfer of learning refers to learning that can be applied to situations that are different from the training situation, which is used as a criterion of training system effectiveness. Importantly, recent research has also shown a positive correlation between SE and user acceptance of VE for learning (Theng *et al.* 2007), and that users with higher levels of SE also achieved a higher learning performance in computer training than individuals with lower SE (Jawahar and Elango 2001, Johnson 2005). Additionally, Marakas *et al.* (1998, p.157) argued that computer SE is a major factor that influences computer performance in challenging skill acquisition situations, due to 'complex mechanisms and relationships that result in increased levels of performance relating to changes in CSE [computer self-efficacy]'. Thus, more research is needed towards this end to better understand the relationship between SE and performance. We hypothesised that in a VE training system:

H1: Self-efficacy will have a positive effect on performance task outcome.

2.2. *Perceived system efficacy*

Studies in information system (IS) research have confirmed positive relationships between SE and perceived behavioural control; and between perceived ease of use of information technology and the perceived usefulness of the technology (Venkatesh 2000, Thompson *et al.* 2008). Compeau and Higgins (1995a) found that the higher the individual's computer SE, the higher his/her affect (or liking) of computer use. Similarly, research (Hill *et al.* 1987, Venkatesh and Davis 1996) found that individuals with a high level of SE have been shown to be more willing to accept and use information technologies. These results suggest SE influences user behaviour and user perceptions of usability of the technology, as well as perceived usefulness of the technology. In our study, we were specifically interested in user perceptions of VE system efficacy. We defined PVE as the extent to which the learning activity required in using a specific VE system is perceived to be effective, efficient and enjoyable. This definition accommodates the users' perceived quality of knowledge transfer afforded by the VE, which can be grouped into three specific measurement focuses: perceived cognitive & learning quality (PCLq); perceived interaction and learning quality (PILq); and perceived system & user interface quality (PSUIq).

Past research (Theng *et al.* 2007, p. 735) refers to quality as output or information produced by the system and defines perceived system quality as the 'perception of how well the system performs tasks that match with job goals'. In VEs the quality of information output and displays are associated with the design features of various I/O devices. Usability of these devices in terms of ease of use, ease of learning and satisfaction are considered as important criteria for usability evaluation (Sutcliffe and Kaur 2000, Stanney 2002, Stanney *et al.*, 2003). In addition, other studies (Moreno and Mayer 2007, Seth *et al.* 2008) have looked into the cognitive aspect of system design and provided theoretical reasoning for cognitive influences on VE efficacy. Although they did not suggest practical measurement methods or tools to access such impact, how well a VE design supports cognitive learning from users' view point is considered an important user preference factor. Furthermore, human–computer interaction and user experience research into VEs (Usoh *et al.* 2000, Whalen *et al.* 2003, Lin 2007) has claimed that it is important to gather users' subjective impressions of their interactions and learning experiences as a way to quantify the design effectiveness of VEs. In particular, when learning materials or data are presented in non-traditional and interactive graphical forms, this may allow for 3D colour graphics and animations that match user interests and increase learning motivation (Saddik *et al.* 2008). Therefore, based on this research that provided a sound theoretical framework, a total of nine parameters (Figure 1) have been identified and associated with PCLq, PILq and PSUIq measurement focuses.

From a theoretical aspect, PCLq is concerned with how users perceive the quality of knowledge transfer evoked by the VE; PILq is concerned with how users perceive their level of interaction with a VE that enabled them to learn effectively; and PSUIq is concerned with how users perceive the effectiveness of a VE system and user interface to enable them to learn effectively. On a practical level, we intended to find out what influences each specific factor under these three main sub-constructs of user PVE on task performance as well as memory (see section 2.5). Past research indicates that SE has a positive effect on user perceived ease of use of a computer system after training (Yi and Im 2004). In the endeavour to understand the relationship between the multimodal information used to quantify VE efficacy, the intention was to find out if user SE has an effect on user perceptions of VE efficacy. Therefore, it was hypothesised that:

> H2: Self-efficacy will have a positive effect on perceived virtual environment efficacy.

2.3. *Performance*

Typically, the performance of a human–computer system is measured through user performance of specific design tasks. Common task performance measures of VEs include time on task and numbers of errors (Nash *et al.* 2000). In the area of a haptically enabled virtual training system (Boulanger *et al.* 2006, Bhatti *et al.* 2008), effectiveness and efficiency of such systems are assessed according to either the technical performance of the haptic interface, or the visual, audio and haptic feedback perceived by users. Measuring the technical performance of the haptic interface often requires algorithm validation and comparison based on rendering realism, whereas measuring various systems feedback perceived by human users comprises methods for the psychophysical evaluation of haptic user interfaces (Samur *et al.* 2007). Many human factor studies have been applied to both assess the performance of haptic user interfaces and user perceived system feedback in sensory-motor control tasks (Ricciardi *et al.* 2009, Sutter *et al.* 2011). In this study, performance was measured objectively on participant behaviour or real-time task performance and subjectively on user perceptions of a virtual training environment (see sections 3.1.1 and 3.1.2). Given that SE has a clear impact on behaviour, such as skill-acquisition (Compeau and Higgins 1995b, Yi and Davis 2003, Hasan 2008), and acceptance of VE technology (Theng *et al.* 2007), and that user perceptions and attitudes are indicators of performance (North *et al.* 2001), studies exploring SE and perceptions of system efficacy on performance are rare. Therefore, it was worthwhile to explore the hypothesised relationship as shown in H1 and H2.

2.4. *Memory*

Performance memory tests of recognition and recall on learnt skills after training have been used as indicators of the effectiveness of a training system (Lin 2004, Hasan 2008). Recognition refers to the understanding of the meaning of the object or environment, and recall is the remembrance of a procedure or an event that occurred in the past (Ryu and Monk 2009). Past research suggested that when participants report their knowledge about a virtual learning environment and the concepts or skills being taught in a memory test, this reflects the effectiveness of the pedagogical aspect of the VE design (Mantovani 2001). In addition, the degree to which a user memorises the features in a VE was also found to be indicative of a subjective sense of 'presence' (Lin 2004). Memory structure of a VE may include the following dimensions – types, shapes, colours, relative locations, relative sizes and event sequences (Usoh *et al.* 2000, Lin *et al.* 2002). Others (Sutcliffe 2003) have claimed that memory test results may reveal potential usability problems in a VE, such as the degree of a subjective sense of 'presence' or 'information quality' and that 'gaps in users' memory, when compared with a gold standard of the information content, point towards presentation problems. A specifically designed memory test questionnaire (MTQ) has been used to aid assessment of the engagement and immersion of a user experience in VE (Lin 2007). Although performance memory tests of recognition and recall on learnt skills after training have been used as indicators of the effectiveness of a training system (Hasan 2008), or indicators of the level of presence (Bowman *et al.* 2004, Lin 2004), one researcher (Hasan 2008) has acknowledged the limitation of using comprehension testing rather than actual task performance in measuring learning. Thus, there was a call for research that used actual task performances to examine the effectiveness of computer technologies. To overcome this limitation, we incorporated both performance of memory tests and tasks in the investigation of VE efficacy. Thus, based on the theoretical and empirical studies described above, PVE is likely to have positive effects on task performance, and objective task performance is likely to correlate with an objective measure of performance memory test. Hence, the following two hypotheses are presented.

H3: Perceived VE efficacy will have a positive effect on task outcome.

H4: Performance task outcome will be positively related to recognition and recall in a performance memory test.

As the above indicates, research has shown strong evidence for a direct relationship between performance and memory (see sections 2.3 and 2.4), and between SE and performance in training settings (see sections 2.1 and 2.4). Research has also demonstrated positive and significant relationships between an individual's affect and attitude as well as performance (see sections 2.2 and 2.3). Importantly, it has been suggested that user attitudes (e.g. affect or liking) towards computers are key indicators of performance in computer-mediated learning (North *et al.* 2001). Also Compeau and Higgins (1995a) found that an individual's computer SE shares a significantly positive relationship with the user affect of computer use, which partially indicates perceived computer system efficacy. Therefore, it is acceptable to hypothesise an interplay between user perceived VE system efficacy and task performance, and between SE and performance. Moreover, as research has shown that performance and memory are positively correlated, it is therefore acceptable to hypothesise that task performance can influence memory, and can mediate the effects of SE and PVE. However, there is a lack of studies to demonstrate any direct relationship between memory and users perceptions. Therefore, we hypothesise that:

H5: No direct relationship will be found between perceived VE efficacy and memory.

H6: No direct relationship will be found between self-efficacy and memory.

Although there is no direct (significant) relationship between the variables as H5 and H6 proposed, that is not to say no relationship exists. It is often the direction (positive or negative) of the relationship and the extent of the correlation (weak, moderate or high) between two variables under investigation which contributes to new insights and findings in a research area. For example, our previous study (Jia *et al.* 2009a) has shown performance (TTS) and memory (MMT) are significantly positively related ($r = .609$, $N = 25$, $P < .05$), and user PVE shares a significant and positive relationship with performance (TTS) ($r = .384$, $N = 30$, $P < .05$). Little is known about to what direction and extent user perceptions, i.e. SE and PVE, share a direct relationship with memory (MMT). As Figure 2 illustrates, TTS may mediate the relationship between PVE and MMT, and between SE and MMT. In this study, we intended to further explore the relationship between these variables as suggested by H5 and H6.

3. Method

The validation of the proposed hypothesis was performed by training users in a new version of an object assembly simulator, called a Virtual Training Environment, developed at the Centre for Intelligent System Research (CISR), Deakin University. Seventy-six volunteers with diverse backgrounds and age-levels performed a series of object assembly tasks in the VTE. Out of these 76 participants, 56 were males and 20 were females. Subjects fell into four age groups: 18–24 ($N = 32$), 25–34 ($N = 33$), 33–45 ($N = 8$) and three ($N = 3$) were over 46 years old. The age groups were divided based on commonly accepted and used age ranges for younger, middle-aged and older adults in HCI and IS literature. For instance, a recent study (Charness *et al.* 2005) examined age and hand performance differences in using light pen and mouse, and has selected participants based on age groups of 18–25, 45–55 and 65–75 to represent young, middle-aged and older adults, respectively in pointing tasks. It is also common in the literature to simply define specific age groups based on the focus of the study leading to the uneven age ranges and groups. For example, a study (Mead and Fisk 1998) which reported age-related performance differences in training ATM menu navigation tasks, has defined two age groups as younger (18–30 years) and older adults (64–80 years) in its investigation. In addition, in an evaluation of VR driving simulator, Liu *et al.* (1999) have targeted age groups of 13–35, 36–55 and 56+ in the investigation of age impact on performance. Furthermore, a more recent study (Yang *et al.* 2008) has involved three women and nine men between the ages of 20 and 27 in validating the performance of haptic motor skill training.

Hardware components of the VE training system included a computer workstation including a Sensable Phantom™ haptics device (6DoF), a Head Mounted Display, a 5DT™ dataglove and a 3D mouse. These hardware components were used to provide users with force feedbacks, 3D object perception, and 3D environment manipulation. Employed virtual training environment for the proposed evaluation study is shown in Figure 3. Software components included a user interface that consisted of a series of user menus and a 3D visual model of assembly objects. The task involved selecting, rotating, releasing, inserting and manipulating 3D objects. These tasks required users to utilise aforementioned input devices.

3.1. Measures

3.1.1. Subjective perception measures: self-efficacy sale and perceive VE scale

In our endeavour to identify factors to quantify VE efficacy, we investigated how people interact and learn in a VE (Jia *et al.* 2009a) and explored ways to measure these factors that quantify VE efficacy. Two factors constituting of SE and nine factors for user PVE were identified, and questionnaire-based evaluation method were adopted to assess these affective learning dimensions of a VE.

Seven items for measuring SE were adopted from previously reported research (Compeau and Higgins 1995b, Johnson and Marakas 2000). These specifically designed instruments were used to measure people's perceptions about their abilities to use a computer successfully – SE in computer skill acquisition (Hasan 2008). Other SE instruments (Wang and Newlin 2002) were also considered to ensure the appropriate design of items.

A large number of items were enclosed in the PVE to ensure a comprehensive evaluation of the VE from a user perspective. To identify items for possible inclusion in the PVE scale, previous studies referring to VE design and evaluation (Preece *et al.* 1994, Brough *et al.* 2007), usability evaluation heuristics (Nielsen 1993, Stanney *et al.* 2003), checklists (Bowman *et al.* 2002, Hix 2002), and superficially designed questionnaires (Kennedy *et al.* 1993, Witmer and Singer 1998, Windell *et al.* 2006) were reviewed. Specifically, Wintmer and Singer's (1998) 'Immersion Questionnaire' (IQ) and 'Presence Questionnaire' (PQ), Lin's (2004) 'Enjoyment, Engagement and Immersion 'scale were considered while designing items associated with the 'Immersion and Presence' parameter. The 'NASA-Task Load Index' (TLX) (Moroney *et al.* 1992) and a self-report instrument (SSI) (Pass *et al.* 2003, Whelan 2007) on 'cognitive load' parameters were used in constructing items for the PVE scale. Moreover, items related to 'usability', 'learnability' and 'satisfaction' parameters were developed with respect to their conceptual meaning (Nielsen 1993, Faulkner 2000, Stanney 2002, Hornbeak 2006). The structure of the questionnaires was not developed on the basis of an 'equal number of items for each parameter', but was coupled with the aims of the SE and perceived efficacy scale, i.e. to capture efficacy parameters from the users' point of view that were deemed to have an effect on quantifying VE efficacy. In addition, questions were presented in a manner that is easy to follow and readily understandable by users. Moreover, side effects and after effects issues were explored using open-ended questions within the PVE scale. Two VE design experts went through the survey instruments of SE and the PVE, and provided feedback on these items. Thus, the initial content/face validity of the instrument was governed.

In addition, as addressed in section 3.1.1, a large number of items were enclosed in the PVE scale, based on previous research (Nielsen 1993, Kennedy *et al.* 1993, Preece*et et al.* 1994, Witmer and singer 1998, Faulkner 2000, Bowman *et al.* 2002, Hix 2002, Stanney 2002, Stanney *et al.* 2003, Paas *et al.* 2003, Lin 2004, Windell *et al.* 2006, Brough *et al.* 2007, Whelan 2007). We argue that this intensive review of related literature enabled us construct the PVE scale, based on a concrete theoretical and academic reference. However, each of the aforementioned research looks into measures of specific user perceptions parameter(s) for instance, 'cognitive load' (Paas*et et al.* 2003, Whelan 2007), 'immersion and presence' (Slater *et al.* 1996, Wintmer and Singer 1998, Nichols *et al.* 2000) which give limited insight of how a VE training system performance as a whole. Since it is imperative to obtain a solid understanding of the important elements that contribute to effective learning through VEs, this study explicitly addressed a complex set of interrelated factors and extended on previously established items to ensure a comprehensive evaluation of VE efficacy based on user feedback.

3.1.2. Objective performance measures: task performance and memory test

Even though subjective measures are becoming an increasingly important tool in system evaluations (Rubio *et al.* 2004), in practice a VE is typically measured objectively on user task performance. Common task performance measures of VEs include time on task, speeds of completion and numbers of errors (Nash *et al.* 2000). Additionally, having computer events drive recordings of all the experiments details, which allows for the incorporation of more accurate performance evaluation of the VE, is also used widely in usability evaluation (Lindgaard 1994, Sutcliffe 2003, Tesfazgi 2003). For example, Lindgaard (1994) explained a good-time logging tool, as a technique of user behaviour observation, allows for quick gathering of data electronically, which is transferable between experimental programmes and systems, aids researchers running parallel sessions on multiple computer stations, and records data from many test users simultaneously in system evaluations. Importantly, objective measurements of the efficacy criteria, such as learnability, can be achieved using electronic data logging during user testing sessions, where completion time for a specific task by a specific set of users; the number of errors per task; and the time spent on using documentation, specific user menus or the help function, can be recorded dynamically (Preece *et al.* 1994). An automatically generated log file that can track performance data, such as task completion rates, time on task, error rates, the number of practices before approaching evaluation tasks, etc. provides objective measures of individual performance (Crellin *et al.* 1990). An electronic data logging device was programmed in this study to accurately and objectively record user task performance in the VE training system. Time and event logs were used to measure user reaction time, time spent completing a particular task, time from committing an error to recovering from it, and amount of task progress during a fixed period of time. This actual performance was also referred to as 'performance usability' (Salzman *et al.* 1999) to distinguish from users 'subjective usability' – user preference or perceptions of an interface. Moreover, the data logging tool enabled the researcher to objectively measure the length of time for the user to successfully perform a benchmark task the first time that the use encountered the VE training system, and the numbers of training sessions this user required to achieve an acceptable performance level indicative of VE learnability. In short, in the current study, users' skill-based learning was assessed through a training test of seven object assembly tasks. Task performance (TTS) was recorded through automatically logging of information of 'time on task' and 'accuracy'.

Importantly, as the performance results were strongly dependent on the tasks to be performed, in the task design, both static objects, i.e. those physically constrained to move within the prescribed limits of the VE, and dynamic objects, i.e. those without any constraints being placed upon their special behaviour (Vince 1995), were included. For example, the car cockpit is a static object that users cannot move around whereas other objects, such as the radio box, screw driver, stereo and power connector are dynamic objects that users could manipulate and manoeuvre through the VE. It is equally important to design tasks that require effective utilisation of a variety of unique I/O devices, at the same time minimum usability flaws in the safety-critical assembly situation, from human factors perspective (Burnett *et al.* 2011, Kostaras and Xenos 2011). Therefore, the display of tasks (assembly objects) and UI components were presented in a manner that is direct, easy to comprehend and institutive.

Seven objects assembly tasks with various levels of difficulty were embedded in the VE training system. The task involved selecting, rotating, releasing, inserting and manipulating 3D objects. These tasks required users to utilise a data glove, a haptics device, a 3D mouse and a head-mounted display (HMD). Higher performance requires users to comprehend assembly sequences and recognise correct objects for specific task procedure as well as utilise various VE input and output devices to achieve learning. Design of the tasks was based on field observation of automotive assembly production line.

Moreover, a performance memory test was also used to assess user cognitive learning. In this study, a memory test was designed to measure cognitive learning in VEs objectively; therefore, questions that did not directly measure an experience, such as 'presence', were treated as perceptually based attributes and measured in specifically designed perception measurement scales (section 3.1.1). Rather, questions in the memory test (MMT) addressed each dimension of the subjects' VE memory structure (e.g. tool used, shape and size) and training procedure (e.g. task sequence, tool support). Sample questions illustrated in Figure 4.

Users were assessed based on their accuracy or recall and the amount of knowledge they learnt from the VE training experience. Users who were able to recall more learning content by answering correctly more memory test questions received a higher performance memory score. Table 2 illustrates the question structure and question score of the retentive/memory test.

Each question was given equal weight, apart from the multiple-choice question Q4, which assigned 25 for the correct two choices of the question. A full score was assigned when the respondents correctly selected the two choices. If only one correct choice was selected, a 12.5 score was assigned. No penalty was given for the wrong choice made, if one correct choice was made. Users were assessed based on their accuracy or recall and the amount of knowledge they had learnt through the VE training experience. Participants who were able to recall more learning content by answering correctly more memory test questions received a higher performance memory score. Similarly, the VE design expert went through the retentive test (MMT) design and provided suggestions on both the content and wording of the questions to ensure they are readily understood by users (with or without technique knowledge of VE). Both learning tests were graded on a 100 scale according to the task difficulty. By focusing on questions related to VE structure and characteristics, the user may reveal his/her spatial awareness, sense of presence and attention on VE. Importantly, both task performance and memory test were reported to reflect the effectiveness of VE system design (Lin 2004).

Many studies have looked into the approaches of measuring VE system efficacy objectively or subjectively. It is likely that system design (e.g. usability, learnability and interactivity) that enhances perceptions (positive affect) might also impact the level of

Figure 4. Sample items in Memory Test questionnaire.

Table 2. Retentive test (memory) question structure and score.

Test questions	Question design	Question weight/score (out of 100)	Question type
Q1	Recognition & recall: *tool used & task sequence*	15	Single-choice
Q2	Recognition & recall: *tool used, tool shape*	15	Single-choice
Q3	Recognition & recall: *tool used, tool shape*	15	Single-choice
Q4	Recognition & recall: *tool used, tool shape*	25 (12.5 + 12.5)	Multiple-choice
Q5	Recognition & recall: *task sequence*	15	Single-choice
Q6	Recognition & recall: *tool support*	15	Single-choice

Note: For the multiple question Q4, 12.5 score was assigned for each correct choice of the two answers.

performance and memory. Additionally, previous research (Lin 2004) suggests task performance may not be a good indicator for the assessment of users' positive affect such as perceptions of enjoyment and satisfaction, but it is essential to be measured to account for the efficacy of VE training systems. Also SE should be a crucial aspect to study since it could be particularly of interest to the VE-based education and training applications. It is therefore important to study both of the user perceptions, i.e. SE and PVE, and the performance, i.e. TTS and memory, i.e. MMT, together when investigating the methods of using multimodal information/measures to enhance design efficacy of a VE.

Overall, four types of outcomes, i.e. SE, PVE, performance, and memory were identified as important indicators of effectiveness of VE systems design. Collection and synthesis of these outcomes enabled us to obtain empirically established data for assessing VE training effectiveness. Importantly, they allowed us to measure cognitive, skill-based and affective aspect of learning (Kraiger *et al.* 1993) that are important in training evaluation. As the research model (Figure 1) exhibits, cognitive learning outcomes were assessed via memory test (MMT), skill-based learning outcomes were accessed via performance training test (TTS) and affective learning outcomes were accessed via two user perception measures, i.e. SE and PVE. Through the

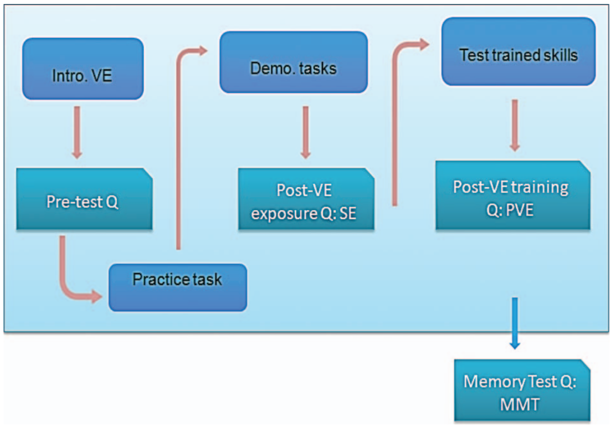

Figure 5. Experiment sequence.

development of the model we hope to discover the approach of best quantifying the VE efficacy as a whole. The objective of this research was to gain a deeper understanding of those relationships among performance, perceptions and memory. As already explained, performance (section 2.3), perception (sections 2.1 and 2.2), and memory (section 2.4) have been considered as crucial aspects of system efficacy. While studying VE efficacy, we consider TTS, SE, PVE, MMT should be assessed simultaneously, as numerous inter-correlations among TTS, SE PVE and MMT could exist.

3.2. Procedure

The study took place in a lab environment. Figure 5 presents the sequence of activities during the experiment.

Upon entering the experimental environment, each participant was given a brief introduction about the purpose of the experiment, the VE training system, the experimental procedure, benefits, possible risks and their rights. A Pre-training questionnaire (Pre-test Q) which gathers their demographic information was then filled out by the participant.

Each participant was then given a brief introduction to the system and performed a simple object assembly task, which served as a pre-test of the participant's ability to interact with, control and use the various VE system control devices. A post-VE exposure questionnaire or SE was then filled out. *SE* was used by participants to subjectively assess one's own capability of performing object assembly tasks in the VE. On a 10-point semantic differential rating scale (from 0 to 100 with 0 being the lowest rating), participants rated their capability to perform a training test with similar types of tasks in terms of accuracy, efficiency and effectiveness. Sample items on the SE include, 'please estimate the accuracy in which you will complete the training test' and 'please indicate the test score that you expect to receive based on accuracy, efficiency and effectiveness'. User level of confidence was rated on the three criteria included in the questionnaire, with the aim to increase the accuracy of data collected. Higher ratings are considered to indicate a more positive belief of SE.

Afterwards, a training test was introduced to the participant. All participants were free to ask questions at any stage during training, related to their training tasks or about the VE training system prior to commencing the final session. In the final session, participants were expected to accomplish seven object assembly tasks within 15 minutes. During this session, the system automatically logged participant's performance on each task (e.g. time on task and error rate). The system also logged the number of attempts made by the user to access the 'help' function provided by the user interface, as well as the restart functionality for any particular sub-task. On completion, a post-test questionnaire or PVE designed to collect the participants' perceptions of the system efficacy, was introduced. PVE was used to measure the individual's beliefs in the effectiveness of the VE to assist them in learning the object assembly tasks. A seven-point Likert scale was used to gather participants' rating for each item, ranging from 1 (strongly disagree) to 7 (strongly agree). Sample statements included '*I was able to focus my attention on learning assembly procedures rather than the input control tools(e.g. haptics device)*'; '*the input control tools (e.g. haptics device, data glove and 3D mouse) were comfortable to operate together in unison*'; and '*I have a strong sense of "being there" (sufficiently immersed) in the virtual training environment*'. Higher ratings are considered to indicate a higher perception of the VE efficacy. All items in the PVE which were negatively poled were recorded so that higher values indicated better ratings.

Lastly, a short interview with each participant was carried out after the test about his or her feelings, emotions, perceptions of the training and learning experience. This was to gather a snapshot of the participant's feeling at a time when they had just experienced the virtual training. The entire experiment including the training sessions lasted about 1.30 hours. One month after the experimental test, an MTQ was distributed online to each participant to assess the level of retention. The MTQ was distributed online to the participants and required them recognise VE I/Q devices, assembly parts, tools and the assembly sequence. Their responses were collected via email.

4. Results

This section is broken down into two main areas: dimensionality of user perception measures, and the relationship between multiple measurement methods and outcomes for VE training efficacy. Four statistical analysis methods, i.e. Factor analysis (FA), Cronbach's alpha, Pearson's correlation and Regression analysis were used, where appropriate, to validate user perception measures and to explore the hypothesised relationships among multiple measurement outcomes.

4.1. *Study 1: dimensionality of self-efficacy and users perceived VE efficacy*

FA for assessing the construct validity and Cronbach's alpha for reliability testing were carried out after the data collection to validate the user perception measures, namely SE and PVE scales. FA and Cronbach's alpha are widely used methods for instrument refinement and validation purposes that are often performed interactively (Kelkar *et al*. 2005, Furr and Bacharach 2008). High alpha coefficients indicate high internal consistent variables of an instrument. A type of FA – principle component analysis (PCA) – was conducted in order to identify the underlying dimensions of efficacy criteria as perceived by users; in other words, to discover and summarise pattern of correlations among variables. Since we already developed a large set of items designed to test the constructs of interest in SE and PVE, respectively, the principal component regression analysis can overcome disturbance of the multicollinearity, and help in selecting items that are ideal/appropriate to measure each of the constructs of interest (Liu *et al*., 2003). Moreover, oblimin rotation was chosen in FA because it was expected that different aspects of SE beliefs and of PVE could be inter-correlated. As a form of 'oblique' rotation, oblimin allows correlations between the factors. On the other hand, 'orthogonal' rotation, such as varimax, is the most commonly used rotation technique in FA (Tabachnick and Fidell 2007). However, varimax is best for extracting factors that are uncorrelated, which was not appropriate for this study and was not used as a preferred method for factor rotation.

Specifically, perceived SE dimensions were developed by submitting seven items to a principal components procedure with an oblimin rotation. This analysis yielded three factors with eigenvalues greater than 1.0, explaining 85.30% of the variance within these data. Factor loading of less than 0.3 was used to omit items that did not load on any of the factors (Hasan 2008). Based on the data, all items were successfully loaded on extracted factors that had a loading above 0.3.

The grouping of items provided insights into the interpretation of the two SE factors. Results confirmed the initially established factors constituting SE. As shown in Table 3, four items loaded on factor 1, which explain 68.44% of total variance. Factor I, labelled *self-efficacy estimate*, is reflective of perceptions of ones' capability in performing tasks in a training test effectively. It consists of four items: estimation of performing task accurately, efficiently and effectively (both accurately and efficiently), as well as an estimate of their test score. The *self-efficacy estimate* factor is illustrative of user beliefs of their capability in performing tasks correctly (accuracy), timely (efficiency) and effectively (both correctly and timely). Factor II, labelled c*onfidence of estimation*, illustrates the confidence level of users on their SE estimation. This factor explains 16.86% of the variance within the sample (Table 3). Three items load onto this factor, confidence of accuracy estimation, confidence of

Table 3. Dimensionality of SE factors.

	Factor loadings		
	I	II	Cumulative
SE factors			
Estimation on task performance – accuracy	.93		.68
Estimation on task performance – efficiency	.88		.83
Estimation on task performance – effectiveness	.86		.84
Estimation of task performance score	.85		.85
Confidence of accuracy estimation		−.99	.92
Confidence of efficiency estimation		−.97	.94
Confidence of effectiveness estimation		−.90	.91
Eigenvalues	4.79	1.18	
% of variance – factor	68.44	16.86	
Mean	45.68	13.28	
SD	12.39	3.77	

Note: Factor loading of less than .3 has been omitted.

efficiency estimation and confidence of effectiveness estimation.

In the same manner, FA was performed to explore the dimensionality of PVE factors. All items had a loading greater than .03; therefore, no item was omitted. Moreover, the importance of each factor is assessed by the percent of variance it represents. The mean score for each factor was calculated by taking into account the factor weight with raw response data. A comprised score was then produced for each factor shown in Table 4.

Moreover, the reliability test of Cronbach's Alpha ($\alpha = .920$) showed that the SE scale is highly reliable, compared with the recommended level (.07) of reliability (Pallant 2000, DeVellis 2003). Reliability test results (Cronbach's Alpha) for the three subscales that measure PVE show PCIq ($\alpha = .87$) and PSUIq ($\alpha = .95$) are highly reliable, compared with an acceptable reliability level of .07. PCLq also showed a satisfactory result ($\alpha = .70$). Specifically, high internal consistency of all factors in each subscales were obtained, with high alpha coefficients ranging from .945 to .948 for PSUIq, .848 to .868 for PILq, and .671 to .750 for PCLq.

Initially, through an extensive review of literature and related studies (Jia *et al.* 2009a, 2009b), we identified nine factors for the measurement focus for PVE. Data from this study empirically assess the appropriateness of these factors. Results show that the factor construct of PVE is unchanged. However, some items were grouped under a different factor than the one to which they were originally assigned. This may contribute to the expected inter-correlations among the factors. Based on the FA and the results from this study, we carefully reviewed the contents of the items and factors they were loaded to, and produced the new labels based on the empirically established factors, as shown in Table 5.

Pearson's correlation test was performed and results show (Figure 6) all sub-scales of PVE were positively and significantly related. Specifically, positive relationships were found between PCLq and PILq ($r = .69$, $p < .05$, $N = 75$); between PCLq and PSUIq ($r = .68$, $p < 05$, $N = 75$); and between PILq and PSUIq ($r = .85$, $p < .05$, $N = 75$). These results also show high criteria-related and construct validity (DeVellis 2003).

4.2. *Study 2: relationship between users perceived VE efficacy dimensions, self-efficacy, task performance and memory*

This study has validated two self-created perception measures, SE scale and perceived VE scale that were developed based on related literature in VE evaluation research and practice. Both measures demonstrate high validity (high correlation coefficiency for criteria-related and construct validity in FA) and internal consistency reliability (coefficient alpha). The means, standard deviations, and correlations among the study variables are presented in Table 6. In particular the results demonstrate SE and PVE share a significant and positive relationship. A graphical representation of the correlations among perception measures and

Table 4. Dimensionality of PVE factors.

	PVE factors								
	Cognitive and learning quality (PCLq)			Interaction and learning quality (PILq)			System and user Interface quality (PSUIq)		
	I	II	III	I	II	II	I	II	III
Eigenvalues	3.26	1.72	1.02	5.17	1.47	1.14	13.08	3.21	1.92
% of variance	36.19	19.13	11.41	43.08	12.23	9.46	38.46	9.43	5.65
Mean	37	12	9	37	8	6	33	9	4
SD	9	5	12	9	2	2	9	2	1

Note: Factor loadings of less than .3 have been omitted, and those judged to constitute a factor. Mean score was calculated according to the factor weight, thus reflects the importance of each factor to the measurement variables.

Table 5. Summary of factors labels from PVE scale.

Factors	PCLq	PILq	PSUIq
I	Objective awareness	Engagement and control	Interactive usability
II	Cognitive load	Interactivity	Visualisation usability (graphical user interface)
III	Learning or knowledge transfer	Immersion (and realism)	Feedbacks (user interface and system)

performance measures is illustrated in Figure 7. As can be seen in Table 6, PVE, task performance and memory all have mean scores greater than 70. Only SE has a lower mean score that is close to 60. Specifically, data indicate that users achieved a high performance task outcome after training. 'Moderate to high' performance of recognition and recall on memory tests was also apparent. In addition, users perceive VE environment to be effective the at 'moderate to high' level, similar to their perception of SE, induced by VE design.

Regression analysis has been widely adopted in IS research (DeVellis 2003, Furr and Bacharach 2008), as a tool to assess and validate research models and hypotheses. It has also been utilised in this study, to explore the hypothesised relationships among perception measures and performance measures (as shown in Figure 2). Although there are other statistical analysis methods such as Structural Equation Modelling (SEM) and Partial Least Square (PLS) that can be used to test the research model, the ordinary regression approach is feasible and adequate in the present study. This is mainly because, there are only a few dependent variables (i.e. performance and memory) and they are not significantly redundant, and have a well-understood relationship to the response. In such cases regression analysis can be a good way to turn data into information (Randall 1995, p.1). Another important reason that we utilised regression analysis over PLS or SEM method is because we were not interested

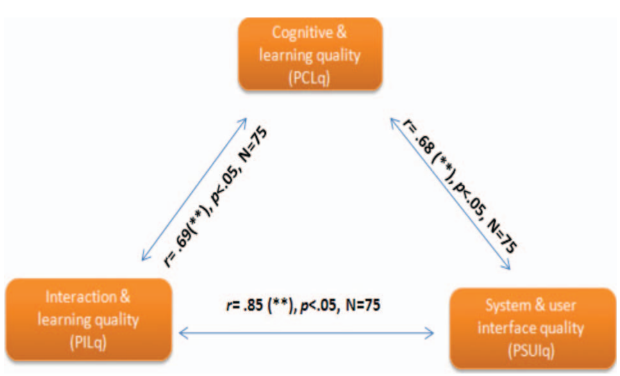

Figure 6. Correlation coefficiency of variables in PVE scale.

in the estimation of latent variables that formal SEM has unique advantages of over PLS (Dykstra 1983, 1985 in Randall 1995), rather our goal was to test the model and understand the relationship between the factors (variables) of interests. To achieve this, we performed a set of separate regression tests for each dependent variable rather than a single multiple regression test with intention to understand the effect and extent of each independent variable (i.e. SE and PVE) on dependent variable (i.e. performance – TTS and memory – MMT).

Analogous to PCA, PLS is a dimensionality reduction technique, which combines features from PCA and multiple linear regression (MLR) to predict a set of dependent variables from a set of independent variables or predictors. This prediction is achieved by extracting a set of orthogonal factors or *latent* variables from the predictors. These latent variables can be used to create displays akin to PCA displays (Abdi 2003). Abdi (2003) also explained that PLS regression is particularly useful when there is a need to predict a set of dependent variables from a (very) large set of independent variables (i.e. predictors), which was not the case in this study. In other words, PLS is effective in the presence of a large number of highly redundant (collinear) factors. More importantly, its emphasis is on the prediction of the responses (variables) and not necessarily on trying to understand the underlying relationship between the variables (Randall 1995). Considering the objective of this study was to better understand the underlying relationship between the variables, regression method was favoured over PLS. Moreover, Randall (1995, p.2) argues that PLS is not usually appropriate for screening out factors as he explains 'if the number of extracted factors is greater than or equal to the rank of the sample factor space, then PLS is equivalent to MLR'.

Furthermore, a past study running analysis of a Technology Acceptance Model (TAM) dataset with intention to address the differences between SEM and regression analysis, found that 'the analyses produced remarkably similar results' (Gefen *et al*. 2000, p.20). Also, despite increased interest and the growing literature of SEM method, 'there is no comprehensive guide for researchers on when a specific form of SEM

Table 6. Means, standard deviations and correlations.

Variable	Mean	SD	*N*	1-SE	2-PVE	3-TTS	4-MMT
1. SE	59	15	76	–	.363 **	−.042	−.499(*)
2. PVE	74	16	76	.363(**)	–	.045	−.342
3. Task performance (TTS)	79	19	75	−.042	.045	–	.033
4. Memory (MMT)	73	20	18	−.499(*)	−.342	.033	–

should be employed' (Gefen *et al*. 2000, p.7). This is in a sharp contrast to the commonly used and widely adopted regression analysis by researchers in the behavioural, social, and educational sciences to assess precision of measurement.

In addition, the present work is heavily influenced by the work of Jawahar and Elango (2001), Hasan (2008), and Thompson *et al*. (2008), and all of these authors have adopted regression analysis as the tool to test theoretical models in their research; therefore, employment of regression analysis becomes an obvious choice in the present study. The results of the regression analyses are presented in Table 7. The regression results show that: (1) a user's perception of SE has a significant effect on PVE ($\beta = 0.413$, $p = 0.000$); (2) SE has no significant effect on task performance ($\beta = -0.052$, $p = 0.659$) and (3) memory ($\beta = -0.231$, $p = 0.342$). Thus, hypotheses 2 – H2 and 6 – H6 were supported and H1 was not supported by the data. With respect to users' perception of VE efficacy (PVE), the results in Table 7 show that (1) PVE has no significant effect on task performance ($\beta = -0.052$, $p = 0.659$) and (2) memory ($\beta = -0.182$, $p = 0.069$). Thus, hypotheses 3 – H3 was supported and 5 – H5 was not supported. Contrary to expectations, the effect of task performance on memory was small and not significant ($\beta = 0.060$, $p = 0.813$). As such, hypothesis 4 – H4 was not supported. Moreover, Table 7 shows that SE explained about 41% of the variance in PVE, and 23% of the variance on the performance memory test. User PVE also explained about 43% of the variance of performance memory test.

The empirical results reported in section 4 via correlation analysis and regression analyses provided solid evidence to either support (accept) or not support (reject) the proposed hypotheses H1, H2, H3 and H4, respectively. The outcomes of 'supported' or 'not supported' presented in Table 7, were all based on the significant value of each hypothesis. Although some hypotheses were 'not supported' or rejected in the study, they still have significance for the research

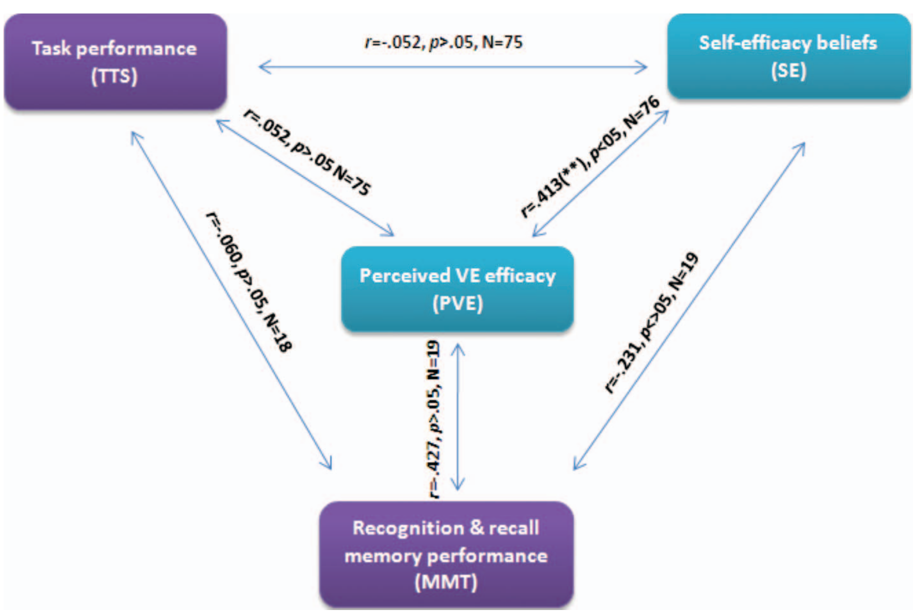

Figure 7. Correlations of perception measures and performance measures.

Table 7. Results of regression testing.

Training outcomes	R^2	β	t	Sig	Hypothesis	Result
Affective and skill: SE → TTS	0.004	−0.063	−0.538	0.592	H1	Not supported
Affective: SE → PVE	0.171	0.413	3.903	0.000	H2	Supported
Affective & Skill: PVE → TTS	0.001	−0.029	−0.244	0.808	H3	Not supported
Skill & Cognitive: TTS → MMT	0.001	0.033	0.108	0.916	H4	Not supported
Affective & Cognitive: PVE → MMT	0.122	−0.349	−1.289	0.222	H5	Supported
Affective & Cognitive: SE → MMT	0.183	−0.428	−0.164	0.127	H6	Supported

field of 3D VEs, as in Yi and Davis (2003) and Corbalan *et al.* (2009) where significant relationship was not found, or the proposed hypotheses were not supported (Chai 2003). For instance, Yi and Davis (2003, p.161) validate an observational learning model in the computer software training and skill acquisition context, and the results showed 'declarative knowledge had no significant effort on delayed performance over and above immediate task performance and post-training SE'. However, as shown in Table 7, all R^2 scores are quite small, which indicates the small variance of each independent variable on the dependent variable. Thus, analysis of combined effect of multiple perception measures on performance and memory may lead to better understanding of the variance of perceptions.

5. Discussion

The purpose of this study was to examine the impact of SE and PVE on task performance and memory, which were considered to be the indicators of VE training effectiveness. A research model positing the relationships of SE, PVE, task performance and memory was developed and tested. Experimental results produced mixed results that partially supported the research model and hypothesised relationships.

As expected, SE and PVE share a significant positive relationship. Accordingly, perceptions of VE efficacy are expected to be higher for individuals with higher SE beliefs than those with lower SE beliefs. This finding is consistent with previous studies that claim SE and user perceptions of computer system effectiveness to be positively associated (Igbaria and Iivari 1995, Jawahar and Elango 2001, Hasan 2008). Importantly, this study extends such an understanding into VE training applications, and demonstrates that SE has a significant impact on PVE. Moreover, the results from this study provided adequate support for two other hypotheses. Specifically, the findings of the study suggest that neither PVE nor SE has a direct and significant effect on memory after training.

In contrast to recent findings, which have shown a significant relationship between SE and performance in computer training, this study shows that SE failed to demonstrate significant effect on task performance in the VE. Participants' unfamiliarity with the VE technology may offer a possible explanation for this finding. Since a VE is not a common technology for novice users (Bowman *et al.* 2004), there is the possibility of developing inaccurate assumptions about their understating of the technology and capability in performing tasks. As Hasan (2008) explained, specific application (where users are familiar with specific training applications, such as VE) SE is usually directly linked to performance in computer training. However, performing VE tasks requires a specific set of skills and cross-domain tasks that it is not possible to capture through SE beliefs. Thus, it is possible that if the VE was more visible and available for trial or if longer training sessions were allowed, then a significant association between SE and task performance may be found. Moreover, a user's PVE has also been shown to have no significant effect on task performance. Research on a TAM (Davis 1993) suggests that user perceptions of technology often shown an association with technology adaption behaviour, whereas, no clear evidence suggests an association between a user perception of technology and their ability to perform tasks using the technology. For example, if users perceive a technology to be useful and easy to use, they are more likely to adopt the technology, i.e. use the technology to assist their work. However, positive perceptions of the technology may not be attributed to how well they are able to use the technology, especially when the technology is not readily assessable by its users, such as VE, and inaccurate judgments between user perception and performance are likely to occur.

Furthermore, unlike previous research (see section 2.3) which measured one or only limited factors of user perception, the measure of user PVE involved comprehensive measures of VE efficacy perceptions from the user's side; therefore, the complexity and completeness of the PVE measure/scale was strong. Thus, the finding may simply be a reflection of the measurement.

Even though user perceptions are often used in VE evaluation, accompanied by task performance, this study focused specific attention on building the connection between the two types of measures. Results did not support the hypothesis, but provided a useful insight into this aspect that can enrich our understanding between perceptually based measure and performance-based measure for VE evaluation. In particular, a related study (Slater *et al.* 1996) showed that one of our established factors, immersion and presence, influence on performance in VEs, and that greater immersion improves task performance. Perceptions of VE efficacy now include a more comprehensive set of factors even though their combined effect on performance may not be very apparent. Furthermore, we found a partial association between task performance and memory.

Previous studies show that usability problems influence task performance outcomes and that users' memory was influenced by usability problems induced by VEs (Sutcliffe *et al.* 2005). This may support the partial connection between the two. Yet the influence of individual differences in terms of retention and age-level differences (David and Fitzgerald 1961, Seufert *et al.* 2009) may contribute to the variance of the association between task performance and memory. VE efficacy score could be produced from a combined analysis of these perception measures and performance measures. Thus, an MLR that takes into account multiple measures may provide a more complete picture of their utility for VE efficacy evaluation, which we intend to perform in our future study. Moreover, other variables, such as prior experience on user skill levels may also be included in the regression analysis. Thus, the effect of user characteristics on performance and memory could be further clarified and enhanced our understanding of the quantification of VE efficacy. Besides, levels of task difficulty have also been shown to impact on performance both in a VE (Riley *et al.* 2004) and a computer environment. Future work could consider exploring this impact on perceptions and performance for quantification of VE efficacy.

One major contribution of this study has been to investigate a complex set of interrelated factors in the relatively new sphere of VEs for training and education. Although many of the factors appear to be important from past research, none of the research has explicitly addressed a set of inter-combination, comprehensive, empirically validated factors to understand how VEs aid complex procedural knowledge and motor skill learning. Specifically, this research has been able to provide empirically established factors for VE efficacy evaluation. A total of 11 factors were derived from two perceptually-based measurement scales, a SE scale and a PVE scale. Consistent with previous studies on VE design and evaluation, it was confirmed that factors such as 'engagement and control', 'cognitive load' and 'interactive usability' are important aspects that account for the effectiveness of a VE design. Importantly, the validation effort also enabled insight into the importance of these empirically established factors. For example, all three 'factor I' in the perceived VE sub-scales are more important than 'factor II' and 'factor III' (Table 5) which account for the variable of perceive VE efficacy. In other words, it is more important that a VE design evoke 'objective awareness', 'engagement and control', and 'interactive usability' (Factor I) than 'learning', 'immersion' and 'feedback' (Factor III).

Another important contribution of this research has been to integrate perceptions of SE and VE efficacy with performance and memory in VE evaluation. For example, even though the study results illustrated no direct impact by PVE on memory ($r = .43$, $p > .95$, $N = 19$), a positive association between the two is apparent. A lack of studies on user perceptions of VE systems and memory made it difficult to compare with result with existing literature. Nevertheless, there is a noticeably increased awareness of the utility of assessing user memory in terms of training and interaction experience. For example, Sutcliffe (*et al.* 2005, p.324) claimed that 'post evaluation memory tests could be usefully incorporated into assessment of VEs as a check on the perceived severity of usability problems'. From past experience and this study, the inclusion of a memory test is critical in VE evaluation. Moreover, this evaluation method also explores users' SE beliefs on the memory, which to the best of my knowledge, has not been studied in VE evaluation. Interestingly, a negative association was found between the two variables ($r = -.231$, $p > .05$, $N = 19$). It shows that people with higher SE beliefs did poorly on the memory test, which contradicted my expectations. Regression analysis further indicated that SE could not be attributed as a key predictor of memory tests results, and only 5% of the variance of memory test was attributable to SE. However, we believe that it is useful to include SE measures in VE evaluation, and that memory tests could be used to check on the perceived SE severity of VE efficacy. According to Gist (1987), it needs to be acknowledged that many factors can influence actual performance in the time span between assessments of SE and performance. Thus, measuring learning performance for an extended period of time after training is a way forward to better understand of perceptions of on performance.

As already addressed, one of the significant contributions of this research has been to produce a set of empirically established factors for VE efficacy evaluation. This has been a critical but missing component of the current literature on VEs, human–VE interaction and virtual training. Importantly, these factors enable VE designers to consider or create new design ideas or solutions. Also most of the studies that apply usability engineering principles to the VE design have primarily involved a small user sample and produced design guidelines or suggestions that were application specific (Bowman *et al.* 2002, Bowman *et al.* 2004). Our research findings are applicable to wider and generic VE applications. This study also confirmed the assumption that users are capable of identifying the critical usability problems of a VE design. A recent study by Sutcliffe *et al.* (2005) involved users, novice observers, and HCI experts in the evaluation of three different types of VE training applications, and demonstrated that a remarkable level of agreement was reached by the three groups of 'evaluator' about the usability problems in the VEs. Thus, exploration of designer–user interaction during training sessions for usability evaluation is an area worthy of further attention in the field of VE training.

6. Conclusion

A comprehensive set of factors has been established through a large empirical study involving 76 participants who underwent training in an object assembly in a VE. Similar studies, interested in VE training effectiveness, have used relatively small samples and used student subjects. This study used larger and more diverse samples to enhance the validity and generalisability of the results. Moreover, the use of multiple measures that provoke multimodal information for quantification of VE training effectiveness also made this study superior to previous ones. For example, not only did the study used actual VE tasks and examine long-term (one to two months after training) effects of training on memory, but also provided empirical evidence of the connections between multimodal information. Extending the prior research on measurement methods and VE training evaluation, this study also demonstrated how user perceptions can affect VE training outcomes.

With respect to VE training, this study took a broader but more complete and systematic approach to evaluate VE efficacy. More specifically, three key learning outcomes (affective, skill-based and cognitive) after training (Kirkpatrick 1987, Kraiger *et al.* 1993) were used to evaluate VE efficacy. Affective learning outcomes were examined with respect to perceptions of SE beliefs and PVE. Skill-based learning outcomes were assessed through a performance test of VE object assembly tasks. Cognitive learning outcomes were examined through a performance memory test. These learning outcomes were considered as indicators of the effectiveness of the VE's design. Through the development of appropriate evaluation methods and an investigation of their associations, this study represents an important attempt to enhance the understanding of the key factors which account for VE efficacy and the impact of these on VE training outcomes.

The results of this study should be useful in the design and administration of VE training programmes. For example, increasing VE experience (e.g. more time on training), enhancing the system and user interface to be more user-friendly (e.g. requiring little or no knowledge of mathematics), and increasing training support (e.g. encouragement during training) are all useful to maximise training outcomes. Increased experience with computers and the implementation of more user-friendly interfaces were also found to be helpful in improving users' SE beliefs (Igbaria and Iivari 1995).

Acknowledgements

This research was fully supported by the Centre for Intelligent Systems Research (CISR) at Deakin University. We wish to thank all the participants who volunteered in the study. We are grateful to the Editor and anonymous referees for a number of critical and helpful comments on an earlier version, which have considerably improved the article.

References

Abdi, H., 2003. Partial least squares (PLS) regression. *In*: M. Lewis-Beck, A. Bryman, and T. Futing, eds. *Encyclopedia of social sciences research methods*. Thousand Oaks, CA: SAGE, 1–7.

Bandura, A., 1997. *Self-efficacy: the exercise of control*. New York: W.H. Freeman.

Bhatti, A., *et al.*, 2008. Haptically enabled interactive virtual reality prototype for general assembly. *In*: M. Jamshidi *et al.,* eds. *Proceedings of the World Automation Congress '08*. Piscataway, NJ: IEEE, 1–6.

Boulanger, P., *et al.*, 2006. *Hapto-audio-visual environments for collaborative training of ophthalmic surgery over optical network*. Paper presented at the IEEE international workshop on haptic audio visual environments and their applications (HAVE), 4–5 November, Ottawa, Canada, 21–26.

Bowman, D.A., Gabbard, J.L., and Hix, D., 2002. A survey of usability evaluation in virtual environments: classification and comparison of methods. *Presence: Teleoprators & Virtual Environments*, 11 (4), 404–424.

Bowman, D.A., *et al.*, 2004. *3D user interfaces: theory and practice*. Boston: Addison-Wesley/Pearson Education.

Brough, J.E., *et al.*, 2007. Towards the development of a virtual environment-based training system for mechanical assembly operations. *Virtual Reality*, 11, 189–206.

Burnett, G., *et al.*, 2011. Designing touchpad user-interfaces for vehicles: which tasks are most suitable? *Behaviour & Information Technology*, 30 (3), 403–414.

Compeau, D.R. and Higgins, C.A., 1995a. Application of social cognitive theory to training for computer skills. *Information Systems Research*, 6 (1), 118–143.

Compeaus, D.R. and Higgins, C.A., 1995b. Computer self-efficacy: development of a measure and initial test. *MIS Quarterly*, 19 (2), 189–211.

Corbalan, G., Kester, L., and van Merrienboer, J.J.G., 2009. Dynamic task selection: effects of feedback and learner control on efficiency and motivation. *Learning and Instruction*, 19, 455–465.

Chai, L., 2003. To have or have not: an examination of feedback, learner control and knowledge type in online learning. *In*: *Proceedings of the 36th Hawaii international conference on system sciences (HICSS'03)*, 6–9 January, Hawaii, USA, 6a.

Charness, N., *et al.*, 2005. Light pen use and practice minimize age and hand performance differences in pointing tasks. *Human Factors*, 46, 373–384.

Crellin, J.M., Horn, T., and Preece, J., 1990. Evaluating evaluation: an empirical examination of novel and conventional usability evaluation methods. *In: Human computer interaction: interact 90*. Cambridge, UK. 329–336.

Davis, F.D., 1993. User acceptance of information technology: system characteristics, user perceptions and behavioural impacts. *International Journal of Man-Machine Studies*, 38, 475–487.

David, P. and Fitzgerald, D., 1961. Meaningful learning and retention: intrapersonal cognitive variables. *Review of Educational Research*, 31 (5), 500–510.

DeVellis, R.F., 2003. *Scale development: theory and applications*. Thousand Oaks: Sage Publications.

Faulkner, X. 2000. *Usability engineering*. Basingstoke: Macmillan.

Furr, R.M. and Bacharach, R., 2008. *Psychometrics: an introduction*. Thousand Oaks: Sage Publications.

Gist, M.E., 1987. Self-efficacy: implications for organizational behaviour and human resource management. *Academy of Management Review*, 12 (3), 472–485.

Gist, M.E., 1997. Training design and pedagogy: implications for skill acquisition, maintenance, and generalization. *In*: M.A. Quinones and A. Ehrenstein, eds. *Training for a rapidly changing workplace: applications of psychological research*. Washington, DC: American Psychological Association, 201–222.

Gefen, D., Straub, D., and Boudreau, M.-C., 2000. Structural equation modeling and regression: guidelines for research practice. *Communications of the Association for Information Systems*, 4 (7), 1–77.

Hasan, B., 2008. The impact of multilevel computer self-efficacy on effectiveness of computer training. *In*: S. Clarke, ed. *End user computing challenges and technologies: emerging tools and applications*. Hershey, PA: Information Science Reference, 33–47.

Hill, T., Smith, N.D., and Mann, M.F., 1987. Role of self-efficacy expectations in predicting the decision to use advanced technologies: the case of computers. *Journal of Applied Psychology*, 72 (2), 307–313.

Hix, D., 2002. Usability engineering of virtual environments. *In*: K.M. Stanney, ed. *Handbook of virtual environments: design, implementation and applications*. Mahwah: Lawrence Erlbaum Associates, 681–699.

Igbaria, M. and Iivari, J., 1995. The effects of self-efficacy on computer usage. *OMEGA International Journal of Management Science*, 23 (6), 587–605.

Jawahar, I.M. and Elango, B., 2001. The effects of attitudes, goal setting and self-efficacy on end user performance. *Journal of End User Computing*, 13 (2), 40–45.

Jia, D., Bhatti, A., and Nahavandi, S., 2009b. Design and evaluation of a haptically enable virtual environment for object assembly training. *The haptic audio-visual environments and games (HAVE09)*, 7–8 November. Piscataway, NJ: IEEE Computer Society: 75–80.

Jia, D., *et al*., 2009a. User-centred evaluation of a virtual environment training system: utility of user perception measures. *In*: R. Shumaker, ed. *Virtual and mixed reality*. Verlag Berline Heidelberg: Springer. LNCS 5622, 196–205.

Johnson, R.D., 2005. An empirical investigation of sources of application-specific computer-self-efficacy and mediators of the efficacy-performance relationship. *International Journal of Human–Computer Studies*, 62, 737–758.

Johnson, R.D. and Marakas, G.M., 2000. The role of behavioural modeling in computer skills acquisition: toward refinement of the model. *Information Systems Research*, 11 (4), 402–417.

Kelkar, K., *et al*., 2005. The added usefulness of process measures over performance measures in interface design. *International Journal of Human-Computer Interaction*, 18, 1–18.

Kennedy, R.S., *et al*., 1993. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology*, 3 (3), 203–220.

Kirkpatrick, D.L., 1959. Techniques for evaluating training programs. *Journal of the American Society of Training Directors*, 13 (11–12), 3–26.

Kirkpatrick, D.L., 1987. Evaluation of training. *In*: R.L. Craig, ed. *Training and development handbook: a guide to human resource development*. New York: McGraw-Hill, 301–319.

Kostaras, N. and Xenos, M., 2011. A study on how usability flaws in GUI design increase mouse movements and consequently may affect users' health. *Behaviour & Information Technology*, 30 (3), 425–436.

Kraiger, K., Ford, J.K., and Salas, E., 1993. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78 (2), 311–328.

Lin, J.J.W., *et al*., 2002. Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment. *In: The IEEE virtual reality 2002,* 24–28 March, Orlando, Florida, USA. New York, NY: IEEE Computer Society, 164–171.

Lin, J.-W.J., 2004. *Enhancement of user-experiences in immersive virtual environments that employ wide-field displays*. Thesis (PhD). University of Washington.

Lin, J.-W.J., 2007. User experience modeling and enhancement for virtual environment that employ wide-field displays. *In*: V.G. Duffy, ed. *Digital human modeling*. Berlin/Heidelberg: Springer, 423–433.

Lindgaard, G., 1994. *Usability testing and system evaluation*. London: Chapman & Hall.

Liu, L., Watson, B., and Miyazaki, M., 1999. VR for the elderly: quantitative and qualitative differences in performance with a driving simulator. *Cyberpsychology & Behavior*, 2 (6), 567–576.

Liu, R.X., *et al*., 2003. Principal component regression analysis with SPSS. *Compute Methods Programs Biomed*, 71 (2), 141–147.

Mantovani, F., 2001. VR learning: potential and challenges for the use of 3D environments in education and training. *In:* R. Giuseppe and G. Carlo, eds. *Towards cyber psychology: mind, cognitions and society in the internet age*, Amsterdam: IOS Press, 207–226.

Marakas, G.M., Yi, M.Y., and Johnson, R.D., 1998. The multilevel and multifaceted character of computer self-efficacy: toward clarification of the construct and an integrative framework for research. *Information Systems Research*, 9 (2), 126–163.

Marcolin, B.L., *et al*., 2000. Assessing user competence: conceptualization and measurement. *Information Systems Research*, 11 (1), 37–60.

Mead, S.E. and Fisk, A.D., 1998. Measuring skill acquisition and retention with an ATM simulator: the need for age-specific training. *Human Factors*, 40, 516–523.

Moreno, R. and Mayer, R.E., 2007. Interactive multimodal learning environments. *Educational Psychology Review*, 19, 309–326.

Moroney, W.F., *et al*., 1992. A comparison of two scoring procedures with the NASA TASK LOAD INDEX in a simulated flight task. *In*: *IEEE national aerospace and electronics conference (NAECON)*, 18–22 May. New York, NY: IEEE Computer Society, 734–740.

Nash, E.B., *et al*., 2000. A review of presence and performance in virtual environments. *International Journal of Human–Computer Interaction*, 12 (1), 1–41.

Nichols, S., Haldane, C., and Wilson, J.R., 2000. Measurement of presence and its consequences in virtual environments. *International Journal of Human–Computer Studies*, 52, 471–491.

Nielsen, J., 1993. *Usability engineering*. San Francisco: Morgan Kaufmann Publishers, Inc.

North, M., *et al*., 2001. Virtual reality and transfer of learning. *In*: J.S. Michael, J.K. Richard, S. Gavriel, and H. Don, eds. *Usability evaluation and interface design: cognitive engineering, intelligent agents and virtual reality*. Vol. 1. London: Lawrence Erlbaum Associates, 634–638.

Ostroff, C., 1991. Training effectiveness measures and scoring schemes: a comparison. *Personal Psychology*, 44 (2), 353–374.

Paas, F., *et al*., 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38 (1), 63–71.

Pallant, J., 2000. Development and validation of a scale to measure perceived control of internal states. *Journal of Personality Assessment*, 75, 308–337.

Popovici, D.M. and Marhan, A.M., 2008. Virtual reality-based environments for learning and training. *In*: D. Talaba and A. Amditis, eds. *Product engineering: tools and methods based on virtual reality*. Dordrecht, Netherlands: Springer, 123–142.

Preece, J., *et al*., 1994. *Human–computer interaction*. Wokingham: Addison-Wesley.

Randall, D. Tobias, 1995. An introduction to partial least squares regression. *In*: *SUGI proceedings*, 2–5 April, Orlando, Florida, 1–8.

Ricciardi, S., *et al*., 2009. Dependability issues in visual-haptic interfaces. *Journal of Visual Languages & Computing*, 21 (1), 33–40.

Riley, J.M., Kaber, D.B., and Draper, J.V., 2004. Situation awareness and attention allocation measures for quantifying telepresence experiences in teleoperation. *Human Factors and Ergonomics in Manufacturing*, 14, 51–67.

Ryu, H. and Monk, A., 2009. Interaction unit analysis: a new interaction design framework. *Human–Computer Interaction*, 24 (4), 367–407.

Rubio, S., *et al*., 2004. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology: An International Review*, 53 (1), 61–86.

Saddik, A.E., Rahaman, A.S.M.M., and Hossain, M.A., 2008. Suitability of searching and representing multimedia learning resources in a 3-D virtual gaming environment. *IEEE Transactions on Instrumentation and Measurement*, 57 (9), 1830–1839.

Salzman, M.C., *et al*., 1999. A model for understanding how virtual reality aids complex conceptual learning. *Presence: Teleoperators and Virtual Environments, 8*, 293–316. Special issue on education.

Samur, E., *et al*., 2007. *Generic and systematic evaluation of haptic interfaces based on test beds. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems*, 29 October–2 November, San Diego, CA. Los Alamitos, CA, 2113–2119.

Seth, A., Su, H.-J., and Vance, J.M., 2008. Development of a dual-handed haptic assembly system: SHARP. *Journal of Computing and Information Science in Engineering*, 8 (4), 44–502.

Seufert, T., Schutze, M., and Brunken, R., 2009. Memory characteristics and modality in multimedia learning: an aptitude-treatment-interaction study. *Learning and Instruction*, 19 (1), 28–42.

Slater, M., *et al*., 1996. *Immersion, presence and performance in virtual environments: an experiment using tri-dimensional chess* [online]. University College London. Available from: www.cs.ucl.ac.uk/staff/m.slater/Papers/Chess/index.html [Accessed 22 May 2012].

Stanney, K.M., 2002. *Handbook of virtual environments: design, implementation, and applications*. Mahwah, London: Lawrence Erlbaum Associates.

Stanney, K.M., *et al*., 2003. Usability engineering of virtual environments (VEs): identifying multiple criteria that drive effective VE system design. *International Journal of Human–Computer Studies*, 58, 447–481.

Sue-Chan, C. and Ong, M., 2002. Goal assignment and performance: assessing the mediating roles of goal commitment and self-efficacy and the moderating role of power distance. *Organizational Behavior and Human Decision Processes*, 89, 1140–1161.

Sutcliffe, A., 2003. *Multimedia and virtual reality: designing multisensory user interfaces*. Hillsdale, NJ: Lawrence Erlbaum.

Sutcliffe, A., Gault, B., and Shin, J.E., 2005. Presence, memory and interaction in virtual environments. *International Journal of Human–Computer Studies*, 62 (3), 307–327.

Sutcliffe, A.G. and Kaur, K.D., 2000. Evaluating the usability of virtual reality user interfaces. *Behavior and Information Technology*, 19 (6), 415–426.

Sutter, C., Musseler, J., and Bardos, L., 2011. Effects of sensorimotor transformations with graphical input devices. *Behaviour & Information Technology*, 30 (3), 415–424.

Tabachnick, B.G. and Fidell, L.S., 2007. *Using multivariate statistics*. Boston: Pearson Education.

Tavakoli, M., Patel, R.V., and Moallem, M., 2006. A haptic interface for computer-integrated endoscopic surgery and training. *Virtual Reality*, 9, 160–176.

Tesfazgi, S.H., 2003. *Survey on behavioral observation methods in virtual environments*. (Research Assignment), Delft University of Technology.

Theng, Y.L., *et al*., 2007. Mixed reality systems for learning: a pilot study understanding user perceptions and acceptance. *In*: *The 12th international conference on human–computer interaction*, 22–27 July, Beijing, China. New York: Springer-Verlag, 728–737.

Thompson, R., *et al*., 2008. Intentions to use information technologies: an integrative model. *In*: S. Clarke, ed. *End user computing challenges and technologies: emerging tools and applications*. New York: Information Science Reference, 79–101.

Usoh, M., *et al*., 2000. Using presence questionnaires in reality. *Presence: Teleoperators & Virtual Environments*, 9 (5), 497–503.

Venkatesh, V., 2000. Determinants of perceived ease of use: integrating control, intrinsic motivation and emotion into the technology acceptance model. *Information Systems Research*, 114, 342–365.

Venkatesh, V. and Davis, F.D., 1996. A model of the antecedents of perceived ease of use: development and test. *Decision Sciences*, 27 (3), 451–481.

Vince, J., 1995. *Virtual reality systems*. New Jersey: Pearson Education.

Wang, A.Y. and Newlin, M.H., 2002. Predictors of web-student performance: the role of self-efficacy and reasons for taking an on-line class. *Computers in Human Behavior*, 18, 151–163.

Whalen, T., Noel, S., and Stewart, J., 2003. Measuring the human side of virtual reality. *In*: *The VECIMS'03, 2003 IEEE international symposium on virtual environments, human–computer interfaces and measurement systems*, 27–29 July, Lugano, Switzerland, 8–12.

Whelan, R.R., 2007. Neuroimaging of cognitive load in instructional multimedia. *Educational Research Review*, 1 (1), 1–12.

Willingham, D.B., 1998. A neuropsychological theory of motor skill learning. *Psychological Review*, 105 (3), 558–584.

Windell, D., *et al*., 2006. A comparison of two mental workload instruments in multimedia instruction. *In*: *The human factors and ergonomics society 50th annual meeting*, 16–20 October. Santa Monica, San Francisco: Human Factors and Ergonomics Society, 1764–1768.

Winn, W., 2002. Learning in artificial environments: embodiment, embeddedness and dynamic adaptation. *Technology, Instruction, Cognition & Learning*, 1, 87–114.

Witmer, B.G. and Singer, M.J., 1998. Measuring presence in virtual environments: a presence questionnaire. *Presence*, 7 (3), 225–240.

Yang, X.-D., Bischof, W.F., and Boulanger, P., 2008. Validating the performance of haptic motor skill training. *In*: *Symposium on haptic interfaces for virtual environments and teleoperator systems*, Reno, Nevada, USA. Madrid, Spain, 129–135.

Yi, M.U. and Im, K.S., 2004. Predicting computer task performance: personal goal and self-efficacy. *Journal of Organizational and End User Computing*, 16 (2), 28–37.

Yi, M.Y. and Davis, F.D., 2001. Improving computer training effectiveness for decision technologies: behavior modelling and retention enhancement. *Decision Sciences*, 32 (3), 521–544.

Yoo, Y.H. and Bruns, W., 2005. Motor skill learning with force-feedback in mixed reality. *In*: *Proceedings of the 9th international federation of automatic control (IFAC) world congress, analysis, design, evaluation of human-machine systems*, 3–8 July, Atlanta.

Copyright of Behaviour & Information Technology is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.