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

Dawei Jia, Asim Bhatti, Saeid Nahavandi

Centre for Intelligent Systems Research (CISR), Deakin University
Geelong, Australia
dwj@deakin.edu.au, asim.bhatti@deakin.edu.au, saeid.nahavandi@deakin.edu.au

Correspondence author’s details:
Dawei Jia
dwj@deakin.edu.au
PH: 61 3 522-72548
Fax: 61 3 5227-1046

Abstract

This study developed and tested a research model which examined the impact of user perceptions of self-efficacy 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. 76 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, results of factor analysis demonstrated dimensionality of two user perception measures and produced a set of empirical validated factors underlining the VE efficacy. Results of regression analysis revealed that self-efficacy had 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 perceptions measures and performance measures for assessing VE training systems efficacy. 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 increased attention and interest given to the utilization 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 features of 3D immersion, multisensory cues, an frames of reference, and employs an advanced human-computer interface (including advanced displays) and 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, 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 traditional useability evaluation methods need to be altered to better suits evaluating 3D VEs (Stanney 2002, 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 factors 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 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 program 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, behavior, and results.

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

Similarly, Kraiger, Ford and Salas (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 program. 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 programs. They further suggest that cognitive learning outcomes focus on the dynamic processes of knowledge acquisition, organization, 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, Ford and Salas (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 utilized theories of cognitive, skill-based and affective learning outcomes for training evaluation (Kirkpatrick et al. 1993).

1.2 Training evaluation for quantification of VE efficacy

Kraiger, Ford and Salas (1993) claim that self-report measures, such as self-rating on self-efficacy or perceived system performance capability, are the most appropriate methods for 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 proceduralization 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 and Daniel 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 self-efficacy 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. self-efficacy and perceived VE efficacy, 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 self-efficacy and perceived VE efficacy) are theorized to influence cognitive (memory) and skill-based learning outcomes (task performance). Specifically self-efficacy may affect memory either directly or indirectly through task performance, and a positive relationship will be found between perceived VE efficacy and task performance. In addition, task performance is theorized to influence memory and mediate the effects of self-efficacy and perceived VE efficacy. Finally, the model posits that self-efficacy may also affect perceived VE efficacy. The theoretical rationale for the model draws upon training evaluation research from both within and beyond the computer-based simulation and training domain. The model is specifically intended to apply within the domain of virtual environment procedural tasks training (e.g. object assembly), and is not designed to be generalized 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.

[Figure 1]
2.1 Self-efficacy

Self-efficacy (SE) refers to an individual’s expectancy in his/her capability to organize and execute the behaviours needed to successfully complete a task (Bandura 1997). It has been conceptualized 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 toward 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 recent study (Hasan 2008) that investigated the impact for self-efficacy on the acquisition of computer skills, he found that a significant and positive effect of SE on far-transfer of learning. 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 self-efficacy also achieved a higher learning performance in computer training than individuals with lower self-efficacy (Jawahar and Elango 2001, Johnson 2005). Additionally, Marakas, Yi and Johnson (1998, p.157) argued that computer self-efficacy 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 toward 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 self-efficacy and perceived behavioural control; and between perceived ease of use of information technology and the perceived usefulness of the technology (Venkatesh and Morris 2000; Thompson, Compeau et al. 2008). Compeau and Higgins (1995a) found that the higher the individual’s computer self-efficacy, 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 self-efficacy 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 perceived VE efficacy (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 (Munro et al. 2002, 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 PSU1q 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 perceived VE efficacy on task performance as well as memory (see 2.5). Past research indicates that self-efficacy 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 self-efficacy has an effect on user perceptions of VE efficacy. Therefore, it was hypothesed 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. 2009), 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 (Bleuler 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 Section 3.1.1 and 3.1.2). Given that self-efficacy has a clear impact on behaviour, such as skill-acquisition (Compeau and Higgins 1995, 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 self-efficacy and perceptions of system efficacy on performance are rare. Therefore, it was worthwhile to explore the hypothesized relationship as shown in H1 and H2 above.

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 memorizes the features in a VE were 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 toward presentation problems. A specifically designed memory test questionnaire 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), nevertheless, 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, perceived VE efficacy 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 three hypotheses are presented.

\[ H3: \text{Perceived VE efficacy will have a positive effect on task outcome} \]

\[ H4: \text{Performance task outcome will be positively related to recognition and recall in a performance memory test} \]

As the above indicates, research has shown strong evidence for a direct relationship between performance and memory (see section 2.3 and 2.4), and between self-efficacy and performance in training settings (see section 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 section 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 (1995) found that an individual’s computer SE shares a significant positive relationship with the user affect of computer use, which partially indicates perceived computer system efficacy. Therefore, it is acceptable to hypothesis an interplay between user perceived VE system efficacy and task
performance, and between self-efficacy and performance. Moreover, as research has shown that performance and memory are positively correlated, it is therefore acceptable to hypothesis that task performance can influence memory, and mediates the effects of self-efficacy and perceived VE efficacy. However, there is a lack of studies to demonstrate any direct relationship between memory and users perceptions. Therefore, we hypotheses 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 perceived VE efficacy (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.

[Figure2]

[Table 1]
2 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 to 25, 45 to 55 and 65 to 75 to represent young, middle-aged and older adults 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 to 80 years) in its investigation. In addition, in an evaluation of VR driving simulator, Liu and others (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 3 woman and 9 men between the ages of 20 and 27 in validating the performance of haptic motor skill training.

Hardware components of the virtual environment 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.
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 utilize aforementioned input devices.

[Figure3]

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 self-efficacy and nine factors for user perceived VE efficacy were identified, and questionnaire-based evaluation method were adopted to assess these affective learning dimensions of a VE.

7 items for measuring self-efficacy 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 – self-efficacy in computer skill acquisition (Hasan 2008). Other self-efficacy instruments (Wang and Newlin 2002) were also considered to ensure the appropriate design of items.

A large number of items were enclosed in the perceived VE efficacy scale (PVE) to ensure a comprehensive evaluation of the VE from a user perspective. To identify items for possible inclusion in the Perceived VE efficacy 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 self-efficacy 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 self-efficacy (SE) and the perceived VE efficacy (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 perceived VE efficacy scale (PVE), based on previous research (Nielsen 1993, Kennedy et al. 1993, Preecet 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, Whelan 2007, Brough et al. 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 look into measures of specific user
perceptions parameter(s) for instance, ‘cognitive load’ (Paas 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, Diaz 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 behavior observation, allows for quick gathering of data electronically, which is transferable between experimental programs 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 1993). An automatically generated log file 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. These actual performance 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 user 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 7 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 utilization of variety 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 et al. 2011). Therefore the display of tasks (assembly objects) and UI components were presented in a manner that is direct, easy to comprehend and instructive.
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 utilize a data glove, a haptics device, a 3D mouse and a head-mounted display (HMD). Higher performance requires users to comprehend assembly sequences and recognize correct objects for specific task procedure as well as utilize 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.

[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.

[Table 2]

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, 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 enhance perceptions (positive affect) might also impact the level of 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 self-efficacy should also 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. self-efficacy, perceived VE efficacy, 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. self-efficacy (SE) and perceived VE efficacy (PVE). Through the 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 (section 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.

[Figure 5]

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 then 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 7 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 7-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, a memory test questionnaire (MTQ) was distributed online to each participant to assess the level of retention. The MTQ was distributed online to the participants and required them recognize 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, 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

Factor Analysis (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 self-efficacy and perceived VE efficacy scales. Factor analysis and Cronbach’s alpha are widely used methods for instrument refinement and validation purposes that are often
performed interactively (Kelkar et al. 2005, Furr 2008). High alpha coefficients indicate high internal consistent variables of an instrument. A type of factor Analysis (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 summarize 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 self-efficacy beliefs and of perceived VE efficacy 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 factor analysis (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 self-efficacy dimensions were developed by submitting 7 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.

[Table 3]

The grouping of items provided insights into the interpretation of the two self-efficacy factors. Results confirmed the initially established factors constituting self-efficacy. 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 *confidence of estimation*, illustrates the confidence level of users on their self-efficacy 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 efficiency estimation and confidence of effectiveness estimation.

In the same manner, factor analysis was performed to explore the dimensionality of perceived VE efficacy 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(a).

[Table 4 (a)]

Moreover, the reliability test of Cronbach’s Alpha (α =.920) showed that the Self-efficacy 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 perceived VE efficacy shows PCIq (α =.87) and PSUIq (α=.95) are highly reliable, compared with an acceptable reliability level .07.PCLq also showed a satisfactory result (α=.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 perceived VE efficacy. Data from this study empirically assesses the appropriateness of these factors. Results show that the factor construct of perceived VE efficacy 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 factor analysis 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 4 (b).

[Table 4 (b)]

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).

[Figure 6]

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, self-efficacy 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 coefficient for criteria-related and construct validity in Factor Analysis) and internal consistency reliability (coefficient alpha). The means, standard deviations, and correlations among the study variables are presented in Table 5. In particular the results demonstrate SE and PVE share a
significant and positive relationship. A graphical representation of the correlations among perception measures and performance measures is illustrated in Figure 7. As can be seen in Table 5, perceived VE efficacy, task performance and memory all have mean scores greater than 70. Only self-efficacy 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 at ‘moderate to high’ level, similar to their perception of self-efficacy, induced by VE design.

Table 5

Figure 7

Regression analysis has been widely adopted in Information System (IS) research (DeVellis, 2003, Furr and Bacharach 2008), as a tool to assess and validate research models and hypotheses. It has also been utilized 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 utilized regression analysis over PLS or SEM method is because we were not interested 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 Principle Component Analysis (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 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, therefore 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) data set 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 should be employed” (Gefen et al.
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, 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 6. The regression results show that: (1) a users’ perception of self-efficacy has significant effect on perceived VE efficacy ($\beta=0.413$, $p=0.000$); (2) self-efficacy has no significant effects 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 6 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 6 shows that self-efficacy explained about 41% of the variance in perceived VE efficacy, and 23% of the variance on the performance memory test. User perceived VE efficacy 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 6, 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 field of 3D VEs, as in (Yi and Davis 2003, Corbalan et al. 2009) where significant relationship was not found, or the proposed hypotheses were not supported (Chai 2003). For instance, Yi and Davids (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 self-efficacy”. However, as shown in Table 6, 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 self-efficacy and perceived VE efficacy on task performance and memory, which were considered to be the indicators of VE training effectiveness. A research model positing the relationships of self-efficacy, perceived VE efficacy, task performance and memory was developed and tested. Experimental results produced mixed results that partially supported the research model and hypothesized relationships.

As expected, self-efficacy and perceived VE efficacy share a significant positive relationship. Accordingly, perceptions of VE efficacy are expected to be higher for individuals with higher self-efficacy beliefs than those with lower self-efficacy beliefs. This finding is consistent with previous studies that claim self-efficacy 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 self-efficacy has a significant impact on perceived VE efficacy. Moreover, results from this study provided adequate support for two other hypotheses. Specifically, the findings of the study suggest that neither perceived VE efficacy nor self-efficacy have a direct and significant effect on memory after training.

In contrast to recent findings, which have shown a significant relationship between self-efficacy and performance in computer training, this study shows that self-efficacy 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) self-efficacy 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 self-efficacy 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 self-efficacy and task performance may be found. Moreover, a user’s perceived VE efficacy has also been shown to have no significant effect on task performance. Research on a technology acceptance model (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 perceived VE efficacy (PVE) involved comprehensive measures of VE efficacy perceptions from the users 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 a 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 self-efficacy scale and a perceived VE efficacy 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 4 (b)) 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 self-efficacy and VE efficacy with performance and memory in VE evaluation. For example, even though the study results illustrated no direct impact by perceived VE efficacy 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 \textit{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’ self-efficacy 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 self-efficacy beliefs did poorly on the memory test, which contradicted my expectations. Regression analysis further indicated that self-efficacy could not be attributed as a key predictor of memory tests results, and only 5% of the variance of memory test was attributable to self-efficacy. However, we believe that it is useful to include self-efficacy measures in VE evaluation, and that memory tests could be used to check on the perceived self-efficacy 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 self-efficacy 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 virtual environments (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 and others (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 generalizability 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 self-efficacy beliefs and perceived VE efficacy. 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 programs. 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 maximize training outcomes. Increased experience with computers and the implementation of more user-friendly interfaces were also found to be helpful in improving users’ self-efficacy beliefs (Igbaria and Iivari 1995).

ACKNOWLEDGEMENT

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.
REFERENCE


Chai, L., 2003. To Have or Have Not: An Examination of Feedback, Learner Control and Knowledge Type in Online Learning. The 36th Hawaii International Conference on System Sciences (HICSS'03), Hawaii, USA, pp. 6a.


Liu RX., Kuang J., Gong Q. and Hou XL. 2003, Principal component regression analysis with SPSS. *Compute Methods Programs Biomed*, 71 (2), 141-147.


Figure 1. Conceptual Framework
Figure 2: A graph depicting the relationships among perceptions, performance and memory.
Figure 3. Employed virtual training environment for the proposed evaluation study
Figure 4. Sample items in Memory Test questionnaire

Did the button menu include a help button?

How many seconds did you use to select the tile?

Which device did you use to enhance the output of the interface menu?

Which tool did you use to fix the screws on the device?

Which part was not used in the entire assembly process?

Which parts did you associate after the virtual training process?
Figure 5. Experiment sequence
Figure 6. Correlation coefficient of variables in PVE scale
Figure 7. Correlations of perception measures and performance measures
Table 1 Hypotheses summary

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Self-efficacy will have a positive effect on performance task outcome</td>
</tr>
<tr>
<td>H2</td>
<td>H2: Self-efficacy will have a positive effect on perceived virtual environment efficacy</td>
</tr>
<tr>
<td>H3</td>
<td>H3: Perceived VE efficacy will have a positive effect on task outcome</td>
</tr>
<tr>
<td>H4</td>
<td>H4: Performance task outcome will be positively related to recognition and recall in a performance memory test</td>
</tr>
<tr>
<td>H5</td>
<td>H5: No direct relationship will be found between perceived VE efficacy and memory</td>
</tr>
<tr>
<td>H6</td>
<td>H6: No direct relationship will be found between self-efficacy and memory</td>
</tr>
</tbody>
</table>
Table 2 Retentive test (memory) question structure and score

<table>
<thead>
<tr>
<th>Test questions</th>
<th>Question design</th>
<th>Question weight/score (out of 100)</th>
<th>Question type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Recognition &amp; recall: <em>tool used &amp; task sequence</em></td>
<td>15</td>
<td>Single-choice</td>
</tr>
<tr>
<td>Q2</td>
<td>Recognition &amp; recall: <em>tool used, tool shape</em></td>
<td>15</td>
<td>Single-choice</td>
</tr>
<tr>
<td>Q3</td>
<td>Recognition &amp; recall: <em>tool used, tool shape</em></td>
<td>15</td>
<td>Single-choice</td>
</tr>
<tr>
<td>Q4</td>
<td>Recognition &amp; recall: <em>tool used, tool shape</em></td>
<td>25 (12.5 +12.5)</td>
<td>Multiple-choice</td>
</tr>
<tr>
<td>Q5</td>
<td>Recognition &amp; recall: <em>task sequence</em></td>
<td>15</td>
<td>Single-choice</td>
</tr>
<tr>
<td>Q6</td>
<td>Recognition &amp; recall: <em>tool support</em></td>
<td>15</td>
<td>Single-choice</td>
</tr>
</tbody>
</table>

*for the multiple question Q4, 12.5 score was assigned for each correct choice of the two answers

Table 3 Dimensionality of self-efficacy factors

<table>
<thead>
<tr>
<th>Self-efficacy Factors</th>
<th>Factor Ladings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>Estimation on task performance - accuracy</td>
<td>.93</td>
</tr>
<tr>
<td>Estimation on task performance - efficiency</td>
<td>.88</td>
</tr>
<tr>
<td>Estimation on task performance - effectiveness</td>
<td>.86</td>
</tr>
<tr>
<td>Estimation of task performance score</td>
<td>.85</td>
</tr>
<tr>
<td>Confidence of accuracy estimation</td>
<td></td>
</tr>
<tr>
<td>Confidence of efficiency estimation</td>
<td></td>
</tr>
<tr>
<td>Confidence of effectiveness estimation</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor Ladings</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalues</td>
<td>4.79</td>
<td>1.18</td>
</tr>
<tr>
<td>% of variance - Factor</td>
<td>68.44</td>
<td>16.86</td>
</tr>
<tr>
<td>Mean</td>
<td>45.68</td>
<td>13.28</td>
</tr>
<tr>
<td>SD</td>
<td>12.39</td>
<td>3.77</td>
</tr>
</tbody>
</table>

*Factor loading of less than .3 has been omitted.
Table 4 (a) Dimensionality of perceived VE efficacy factors

<table>
<thead>
<tr>
<th>Perceived VE efficacy (PVE) factors</th>
<th>Cognitive &amp; Learning quality (PCLq)</th>
<th>Interaction &amp; learning quality (PILq)</th>
<th>System &amp; User Interface quality (PSUIq)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>3.26</td>
<td>1.72</td>
<td>1.02</td>
</tr>
<tr>
<td>% of variance</td>
<td>36.19</td>
<td>19.13</td>
<td>11.41</td>
</tr>
<tr>
<td>Mean</td>
<td>37</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>SD</td>
<td>9</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

*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 reflect of the importance of each factor to the measurement variables

Table 4 (b) Summary of factors labels from PVE scale

<table>
<thead>
<tr>
<th>Factors</th>
<th>PCLq</th>
<th>PILq</th>
<th>PSUIq</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Objective awareness</td>
<td>Engagement and control</td>
<td>Interactive usability</td>
<td></td>
</tr>
<tr>
<td>II Cognitive Load</td>
<td>Interactivity</td>
<td>Visualization usability (Graphical user interface)</td>
<td></td>
</tr>
<tr>
<td>III Learning or knowledge transfer</td>
<td>Immersion (&amp; realism)</td>
<td>Feedbacks (user interface and system)</td>
<td></td>
</tr>
</tbody>
</table>
Table 5 Means, standard deviations and correlations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>1-SE</th>
<th>2-PVE</th>
<th>3-TTS</th>
<th>4-MMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Self-efficacy (SE)</td>
<td>59</td>
<td>15</td>
<td>76</td>
<td>--</td>
<td>.363 **</td>
<td>-0.042</td>
<td>-.499(*)</td>
</tr>
<tr>
<td>2. Perceived VE efficacy (PVE)</td>
<td>74</td>
<td>16</td>
<td>76</td>
<td>.363( **)</td>
<td>--</td>
<td>.045</td>
<td>-0.342</td>
</tr>
<tr>
<td>3. Task performance (TTS)</td>
<td>79</td>
<td>19</td>
<td>75</td>
<td>-0.042</td>
<td>.045</td>
<td>--</td>
<td>0.033</td>
</tr>
<tr>
<td>4. Memory (MMT)</td>
<td>73</td>
<td>20</td>
<td>18</td>
<td>-0.499(*)</td>
<td>-0.342</td>
<td>0.033</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 6. Results of regression testing

<table>
<thead>
<tr>
<th>Training outcomes</th>
<th>R²</th>
<th>β</th>
<th>t</th>
<th>Sig</th>
<th>Hypothesis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affective &amp; Skill: SE-&gt;TTS</td>
<td>0.004</td>
<td>-0.063</td>
<td>-0.538</td>
<td>0.592</td>
<td>H1</td>
<td>Not supported</td>
</tr>
<tr>
<td>Affective: SE-&gt;PVE</td>
<td>0.171</td>
<td>0.413</td>
<td>3.903</td>
<td>0.000</td>
<td>H2</td>
<td>Supported</td>
</tr>
<tr>
<td>Affective &amp; Skill: PVE-&gt;TTS</td>
<td>0.001</td>
<td>-0.029</td>
<td>-0.244</td>
<td>0.808</td>
<td>H3</td>
<td>Not supported</td>
</tr>
<tr>
<td>Skill &amp; Cognitive: TTS-&gt;MMT</td>
<td>0.001</td>
<td>0.033</td>
<td>0.108</td>
<td>0.916</td>
<td>H4</td>
<td>Not supported</td>
</tr>
<tr>
<td>Affective &amp; Cognitive: PVE-&gt;MMT</td>
<td>0.122</td>
<td>-0.349</td>
<td>-1.289</td>
<td>0.222</td>
<td>H5</td>
<td>Supported</td>
</tr>
<tr>
<td>Affective &amp; Cognitive: SE-&gt;MMT</td>
<td>0.183</td>
<td>-0.428</td>
<td>-0.164</td>
<td>0.127</td>
<td>H6</td>
<td>Supported</td>
</tr>
</tbody>
</table>