Systematic Evaluation of a Virtual Environment-based Training System

Dawei Jia, Asim Bhatti and Saeid Nahavandi
Centre for Intelligent Systems Research (CISR), Deakin University
dwj@deakin.edu.au; asimbh@deakin.edu.au; nahavand@deakin.edu.au

Abstract: This paper explores the application of three constructs that deemed to be essential to quantify virtual environments (VE) efficacy: cognitive, skill-based, and affective learning outcomes. The authors discuss the implementation of these constructs in a user-centered evaluation of a VE training system. By transforming both the conceptual and operational cohorts for training evaluation the authors illustrate the benefits of the development of a Multi-dimensional User-centred Systematic Training Evaluation (MUSTe) method for quantifying VEs efficacy. Importantly, MUSTe acknowledges the importance of combining holistic and analytical approaches in conducting systematic user-based evaluation. Furthermore, it also emphasizes that quantifying VEs efficacy must reflect the perception and preferences of the users rather than the imposition of efficacy on single measures of task outcome. An empirical study that applied MUSTe evaluation method in quantifying a VE training system efficacy provided valuable evidence of the theoretical construct and content validity of the method.

1. INTRODUCTION

Virtual Environment (VE) is a computer-generated, 3D spatial environment that enables user to interact with, via multiple human sensorial channels. During recent years VE has become a promising tool for training and education [2]. Despite its adaptation for training and fast-paced technological advancements, ways in which to evaluate efficacy of such technology are unclear [18].

It is been argued that the key contributor to such deficiency is due to a marked absence of research on the development of evaluation methods that suits users and activate their role in evaluation. To address the coupling of user-centered design and evaluation, we propose an evaluation methodology that extends and draws on system, task and user performance metrics proposed by Bowman and others [1], with intention that it allows a systematic way of accessing and quantifying VE efficacy for training. Joining a handful of researchers who have endeavored to understand relationship between efficacy of VEs and user interaction and learning experience, we illustrate how cognitive, skill-based and affective learning outcomes [9] can be used and measured to quantify VE efficacy. “Efficacy”, in a broadest sense is defined as “how effective a VE is in assisting users to achieve intended learning outcomes”. The definition of efficacy encompasses both the quality of interaction experience and levels of learning outcomes. This research was grounded in two areas of evaluation research – how people learn in technology-mediated context and training evaluation design and methods. In this paper we report on the development of a training evaluation method for quantifying VEs efficacy and report findings on the validation of this method.

2. RELATED WORK

2.1 How people learn in technology-mediated environment

Some argue that technology plays an important role in shaping interaction process and learning experience [8] [14]. Others believe that learners are active players whom constantly building mental models and ‘making senses’ in technology mediated learning context [17] [13], and technology is merely a ‘tool’ to support learners’ construction of knowledge, structure their learning process and stimulate learner to make maximum use of their cognitive potential. Both view points explain learning from either technology-oriented or learner-oriented perspective, yet overlooked the ‘coupling’ and ‘interactivity’ between users and technology in support learning.

Coupling, as explained by Winn [20] is a mutually influencing dynamic interaction between users/learners and the technology-mediated environment. In his recent work “Embodied and Embedded Cognition” (ECC), Winn emphasizes the bio-directional influences between learner and learning environment, and argues that the actions of learner and environment and consequences of these actions on each other makes the two systems tightly coupled. Interactivity, is another key factor that influence overall user experience and learning in technology-mediated context [14] [10] [11]. For example, user experience of presence, enjoyment and simulator sickness in VE are all co-related with levels of interactivity [10]. Benefits of ensuring coupling and interactivity between learner and technology-mediated learning environment can be evident from empirical studies [10] [11], which shown enhanced learning experience and outcomes. Additionally, better interactivity produces more pleasing, better-controlled interactions in technology-mediated learning environment [12][10].

Ensure learner is tightly coupled to the learning environment require high level of presence, which need
complete attention and total engagement of learner [20] [13]. Moreover, cognitive and affective strategies a learner engage to couple himself/herself to the learning environment reflect his or her perception of the environment and senses of presence, which may lead to differences in performance [6]. More important, Hudlicka and McNeese [5] found that users’ affective and belief states influence a variety of perceptual, cognitive, and motor processes, both in low-level processes and higher-level processes. Therefore, these affective and beliefs strongly influences user task performance in variety of systems, including virtual reality training environments and instructional systems.

2.2 Training evaluation design and methods

Psychologists have long recognised the importance of human cognition and affect in evaluating computer-based learning system. While design features of computer systems influence on cognitive engagement (such attention, comprehension and cognitive load) and affect (such attitude and perception), and people regulate their cognitive and affective strategies constantly when performing tasks in computer-based training systems, in most cases training evaluation fail to account for these factors [9]. Kraiger, Ford and Salas [9] argue that to quantify the success of training programs, systematic collection of data that measure multidimensional learning outcomes are essentials. For example, training evaluation needs to take a construct-oriented approach that measure learning in terms of cognitive, affective and skill-based outcomes.

Others have suggested that motivational and affective factors have significant impact on human learning activities [13] and a dynamic interaction and adaption between learner and his/her environment are formed by regulating cognitive engagement and affect [13]. For example, in an interactive virtual learning environment, affective strategies that couple learner and VE is through engage learner’s sense of “presence”- feelings and believes of being “in” the computer-generated world. In addition, affective and cognitive factors can influence the level of coupling between learner and his or her computer-generated environment when learners actively involving their bodily activities [20]. Despite this, cognitive engagement and affect may also influence skill-based learning outcomes [3] [9].

Moreover, evaluation of VEs in practice often adopts an analytical approach to a single aspect of the VE, this has the advantage in specific diagnosis of system problems [4]. However it treats the system as separate components, and only discovers problems at the low level, thus evaluation of the system as a whole could not be achieved. On the other hand, holistic approach could serve such a purpose, a genetic VE system consists of system components and user interface components, it is surprising to see a lack of research applying systematic approach to the evaluation of VEs. With this in mind we proposed a Multi-dimensional User-centered Systematic Training Evaluation (MUSTe) method for quantifying VEs efficacy. The motivation of this research is to develop a reliable, valid and effective evaluation methodology for quantifying VE efficacy. The main novelties of the MUSTe methodology are as follows: firstly, it considers multi-dimensions cognitive, affective and skill-based learning to be equally important in training evaluation. Secondly it proposes a user-centered evaluation approach; this means the users’ point of view is used in draw evaluative conclusions of the efficacy of VE training systems. Users’ behaviors, feedbacks and preference all shape and influence the evaluation outcome. Finally it allows a systematic evaluation that not only incorporates evaluation parameters in reference to multi-dimensions, but also encompasses specifically designed measurement tools that aim to achieve evaluative conclusion in a systematic and holistic way. In this paper, we further explore the ways to systematically evaluate VEs efficacy.

3. MUSTe EVALUATION METHODOLOGY

3.1 Theoretical construct

Salzman et al [16] suggested that VE’s affordances work with other interaction and learning experience factors that influence the quality of learning outcomes. When presented with learning tasks in a VE system, learners constantly ‘making sense’ and building mental models. During this process, the perception of the sensory information takes place, and attitude and emotions are integrate within this process. This creates an alternate perception in users’ mind that is tightly enveloped with cognitive and affect strategies, resulting in a unique learning experience afford by VE features. For instance, allow mixed modalities to engage human perceptual, cognitive, and communication skills in understanding what is being presented in a virtual world [19]. Pioneers in the field of 3D user interface and VEs have illustrated that features of virtual learning environment do not act in isolation to achieve intended learning objectives [12]. Learning tasks, learner characteristics, learning and interaction experience are all play a role in shaping the learning process and learning outcomes. System design feature influence on learning process and outcomes is presented in figure 1.

Figure 1. VE design features influence on learning process and learning outcomes (based on [16] and [9])

In our endeavor to identify and measure factors to quantify VE efficacy, we investigated how people
interact and learn in a VE and explored ways to measure these factors that quantify VE efficacy [7].

3.2 MUSTe measures

An integral part of the MUSTe is the use of questionnaires designed for users to assess the efficacy of VEs. Questionnaire-based evaluation method has proven to be effective in evaluating 2D interface/application in various domains. However, not many studies dedicate in the design of questionnaire-based evaluation tool for user-based testing in the field of VE. MUSTe adopt questionnaire-based evaluation method for measure user affect dimension of VEs training system is described. Access affective learning outcomes are through post-VE exposure questionnaire and post-VE training questionnaire (see 3.2.4 for experimental procedure) that require users to subjectively rate their perceived interaction and learning experience. Post-VE exposure questionnaire measure users’ believes of self-efficacy, which require users to predict their task outcomes in terms of accuracy, efficiency and effectiveness. Empirical evidence illustrates that attitude direction and strength, and self-efficacy are the key affective learning outcomes that can be collected through self-report measures. Post-VE training test questionnaire contains items that measure attributes of identified efficacy factors. In lie with researchers [15] we believe that thoughtful application of theory to practice should reveal the potential of VE efficacy for learning through conventional assessment, and that a questionnaire instrument can be used to evaluate VE efficacy.

In order to access skill-based learning outcomes, objective measure of user task performance of ‘task completion rate’, ‘time on task’ and ‘error rate’ are recorded through logging file. To access cognitive learning outcomes, a memory test questionnaire was designed to collect data from users on their accuracy or recall and amount of knowledge they learnt from VE. Memory test has been used to aid assessment of engagement and immersion of user experience in VE [10]. By focusing on questions relate to VE structure and characteristics, user may reveal his/her spatial awareness, sense of presence and attention on VE.

In addition, proposed evaluation method is heavily based on the hypothesis that “the efficacy of VE training system is quantified based on three dimensions of learning outcomes - cognitive, skill-based and affective”. It is anticipated that better perception and positive attitude of user interaction and learning experience within the virtual environment will result in higher cognitive, affective and skill-based outcomes, determining the efficacies of the VE systems. Furthermore, due to the obvious effects of users’ prior experience on task performance, it is anticipated that higher learning outcomes are achieved by more experienced users.

3.2.1 Virtual Training Environment (VTE) setup

Hardware components of the VTE include an Intergraph workstation that runs a Sensable P+HANTOM haptics device (6DoF), a Head Mounted Display, and a 3D mouse. These hardware components were used to provide users with force feedbacks, 3D object perception, and 3D environment manipulation. Software components include a user interface that consists of a series of user menu and 3D visual model of assembly objects.

3.2.2 Object Assembly Tasks

The experimental tasks are object assembly of a car cockpit. Each subject was required to perform 7 object assembly tasks in the VTE via 4 main task sequences. Each task included several activities: picking, rotating and releasing object; manipulating 3D environment, and viewing and assemble required objects. Figure 2 displays screenshots of the object assembly tasks each subject practice (a) and tested (b).

3.2.3 Subjects

A total of 30 subjects (4 female and 26 male) participated in this study were recruited from Deakin university. The average age of the participant population fell between 25-34 years. Of these subjects, 7 were very experienced VE users (VE expert-VEP), 11 were experienced with object assembly tasks (Task expert-TEP), and 12 were neither experienced users of VE or experienced in object assembly tasks (Non expert-NEP). This study was approved by the Deakin University Ethics Committee.

3.2.4 Procedure

Upon entering the experimental environment, each subject was asked to complete a pre-test questionnaire. Each subject was then given a brief introduction of the system and performs a simple object assembly task, which serves as a pre-test of subject’s ability to interact with, control and use various VR system control devices (Head-Mounted Display, 3D mouse, Data glove and Haptics device). A self-efficacy questionnaire was then filled out. Afterwards, a training test was presented to each subject, whom has 15 minutes to complete all 7 object assembly tasks in the VR system. A post-test questionnaire was presented to the subject in the experimental environment. At last, an open-ended interview with each subject was carried out right after the test, which was part of the video recording. Two weeks after the experimental test, subjects required to
respond on a memory-test questionnaire that requires them to recall their learning tasks or procedures in the VR training system.

4. RESULTS

VE efficacy was hypothesized to be significantly affected by different levels of prior experience in manipulating 3D objects in gaming or computer environment (LOE3D). As VE efficacy was measured on TTS, SelfEfficacy, PVEefficacy and MMT, it was expected that people with higher level of LOE3D have higher self-efficacy beliefs, achieve better outcome in training test, perceive the VE to be more effective and have higher achievement on the memory test.

4.1 Skill-based learning outcomes: User task performance

On average, subjects achieved task score of 64.67 (out of 100). As shown in Figure 3, mean score for TEP is 78.18, NEP achieved 38.33 and 88.57 for VEP. Overall, VEP and TEP achieved similar levels of object assembly score, with VEP outperform TEP on each assembly task (but task 6), and NEP achieved least task score in object assembly compare with TEP and VEP.

![Figure 3. Skill-based task outcomes](image)

In addition, one-way between groups ANOVA with planned comparisons were performed. At significant level of .05, an examination of the Levene test for homogeneity of variances suggests that this assumption has not been violated ($p > .05$). The F-ratio with an F-probability value less than .05 is significant, suggesting that different type of expertise (VEP, TEP and NEP) does significantly influence task score in VE training system. Post-hoc analyses further shows that the difference lays between NEP and VEP ($p=.000$); and between NET and TEP ($p=.000$). No significant difference found between VEP and TEP on task score.

4.2 Affective learning outcomes: User perception

4.2.1 Self-efficacy

To measure subjects’ beliefs of self-efficacy in the VE training system, subjects’ estimation of accuracy (EstAccuracy), efficiency (EstEfficiency) and effectiveness (EstEffectiveness) were collected. Due to the potential effects of individual difference on estimation, each subject rating on his or her confidence of EstAccuracy (ConfAccEst), confidence of EstEfficiency (ConfEffEst), and confidence of EffEst (ConfEffEst) were also gathered to provide additional assurance for their estimation. To simply the results obtained from this exploratory study, estimation of training test score (EstTTS) was used to represent the outcome of self-efficacy.

![Figure 4. Beliefs of Self-efficacy](image)

Interestingly, mean score for self-efficacy between VEP and TEP differs greatly compare with the similar results for NEP and TEP as shown in Figure 4. However, one-way between ANOVA analysis shows the difference is not significant ($F=1.918$, $p=.170$). Therefore, individuals tend to have similar believes of self-efficacy, regards to their differences in expertise.

4.2.2 Perceived VE efficacy

To quantify perceived VE efficacy, three factors: perceived cognitive learning quality (PCLq), perceived system and interface quality (PSIq), and perceived interaction and learning quality (PILq), were measured based on response from post-VE training test questionnaire. Perceived VE efficacy was generated based on mean score on these three factors.

![Figure 5. User perception of VE efficacy](image)

Mean score of user perception on VE efficacy shows that subjects with different expertise perceive VE to be equally effective: on a 100 scales, mean subjective rating for all user groups greater than 60, and closer to 70 for VEP and TEP. In addition, one-way between ANOVA analysis support the mean result and shows...
that no significant differences found among subjects on perceived VE efficacy ($F=.385, p>.05$).

4.3 Cognitive learning outcomes: Accuracy of recall

As the memory test (MMT) questionnaire distributed 2 weeks after the experiment test, not all subjects respond. Results were based on response from 26 subjects. Of these 26 subjects, 7 were VEP, 9 were TEP and 10 were NEP. On average, subjects achieved 84.80 out of 100, with 74.25 for NEP, 88.89 for TEP and 94.64 for TEP. Moreover, ANOVA analysis shows that there is a significant difference on MMT cross different expertise ($F=7.215, p<.004$). Post Hoc test further reveals that the difference lays between VEP and NEP ($p=.004$), and between NEP and TEP ($p=.030$). No significant difference found between VEP and TEP ($p>.05$).

4.4 Effect of prior experience

VE efficacy was hypothesized to be significantly affected by different level of prior experience in manipulating 3D objects in gaming or computer environment (LOE3D). As VE efficacy was measured on TTS, SelfEfficacy, PV.Efficacy and MMT, it was expected that people with higher level of LOE3D have higher self-efficacy beliefs, achieve better outcome in training test, perceive the VE to be more effective and have higher achievement on the memory test.

To assess the utility of prior experience for explaining task outcome, we used multiple predictors: computer use frequency (CompFreq), computer use history (CompHis), experience of manipulating 3D objects in gaming or computer environment (LOE3D), experience of manipulating 3D objects in VE environment (ExpVE). These were included in a multiple linear regression (MLR) model to predict training test score (TTS). Because of potential effects object assembly skills in real life may have influence on the subjects’ performance in the VE, experience of using electronic tools for object assembly tasks (ExpTool), and perceived level of difficulty of assembly task (PdifTask) were included as predictors in this model. Finally, due to the potential effects of age and gender on training test score, and other response measures, these two variables were included in the model.

In general, the inclusion of these variables in the predictive model of training test score was aimed at avoiding biases in the parameter estimates; CompFreq, CompHis, LOE3D and ExpVE that might have occurred if variance due to prior object assembly skills (ExpTool, PdifTask) or individual differences were not taken into account. However, it is anticipated that there were interrelationships among the variables. With this in mind, standard approach of multiple regression was performed, which allowed us to find out how the multiple predictors combine to influence the training test score. The regression model used to assess the utility of multiple predictors on training test score was structured as shown in equation 1.

\[
TTS = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} + \beta_3 \text{CompFreq} + \beta_4 \text{CompHis} + \beta_5 \text{LOE3D} + \beta_6 \text{ExpVE} + \beta_7 \text{ExpTool} + \beta_8 \text{PdifTask}
\]

Results of the standardized regression coefficients analysis indicated that this regression model predicts training test score well, $F$ (2.404), $p<0.05$. Approximately 48% of the variability in training test score was explained by this model ($R^2=0.478$). The results also show that at the $\alpha=0.05$ level, LOE3D is the most important predictor of training test score ($\beta_{\text{LOE3D}}=0.567, p=0.032$). More important, LOE3D alone, account for 38% of the variance of training test score, $F=17.136, p<0.000$. Surprisingly, of the eight predictors, only subjects’ prior experience of manipulating 3D object in gaming or computer environment contributes significantly ($p$=0.001) to the model. Correlation analysis (1-tailed) also confirms that LOE3D was significantly and positively correlated with training test score, $r=.616, N=30, p<0.000$. In other words, people who are more experienced in manipulating 3D objects in gaming or computer environment tend to achieve higher training test score. In addition, a moderate but significant linear relationship between gender and training test score ($r=0.321, N=30, p=0.042$), and between ExpVE and training test score ($r=.358, N=30, p=.026$) were found. These results show that male tend to outperform than female, and people with more experience in manipulating 3D objects in VE achieved higher training test score. In addition, younger people tend to have more experience of manipulating 3D objects in gaming or computer environment than elder ones, $r=0.508, N=30, p=0.004$.

Table 1. Results of Standardized Regression Coefficients Analysis on Individual Parameter Estimates

<table>
<thead>
<tr>
<th>Variables</th>
<th>df</th>
<th>t Value for $H_0$</th>
<th>Prob &gt; T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.11</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>0.063</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1.27</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>CompFreq</td>
<td>1</td>
<td>0.51</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>CompHis</td>
<td>1</td>
<td>0.693</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>LOE3D</td>
<td>1</td>
<td>2.301</td>
<td>$p &lt; 0.05$</td>
</tr>
<tr>
<td>ExpVE</td>
<td>1</td>
<td>0.403</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>ExpTool</td>
<td>1</td>
<td>0.314</td>
<td>$p &gt; 0.05$</td>
</tr>
<tr>
<td>PdifTask</td>
<td>1</td>
<td>1.130</td>
<td>$p &gt; 0.05$</td>
</tr>
</tbody>
</table>

*significant at the $\alpha = 0.05$ level

5. CONCLUSION

Our results confirm that VE design features can influence user task performance and cognitive learning outcomes regardless of their different levels of expertise. Subjects have achieved similar levels of object assembly in the virtual training environment. Clear effects of subjects’ expertise on task performance were found in this study. Experienced subjects of VE
Nevertheless, there are several factors which need to be considered in evaluating the findings of the present research. First, although the observed differences in task performance were statistically significant, they are small in absolute terms. Secondly, subjects were selected who lacked obvious physical infirmities or disorders, it became apparent during the course of the study that the ability and length of wearing Head-Mounted Display (HMD) were not comparable in terms of stereograpical vision. Differences in these may therefore have contributed to the observed differences in object assembly speed and accuracy, as well as subjects’ self-efficacy believes. Finally, the present study demonstrates a basis for user-centered evaluation in only one domain: efficacy of VE training system of object assembly. It remains to be determined whether these differences hold true across other situation, or are specific to the VE object assembly simulator.

REFERENCES