The stress and workload of virtual reality training: the effects of presence, immersion and flow

Article *in* Ergonomics · March 2016

DOI: 10.1080/00140139.2015.1122234

CITATIONS
116

READS
2,346

4 authors, including:

Stephanie J. Lackey
University of Central Florida
43 PUBLICATIONS **420** CITATIONS

SEE PROFILE

James L Szalma
University of Central Florida
191 PUBLICATIONS **4,848** CITATIONS

SEE PROFILE

Peter A Hancock
University of Central Florida
648 PUBLICATIONS **27,698** CITATIONS

SEE PROFILE

Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: http://www.tandfonline.com/loi/terg20

The stress and workload of virtual reality training: the effects of presence, immersion and flow

S. J. Lackey, J. N. Salcedo, J.L. Szalma & P.A. Hancock

To cite this article: S. J. Lackey, J. N. Salcedo, J.L. Szalma & P.A. Hancock (2016): The stress and workload of virtual reality training: the effects of presence, immersion and flow, Ergonomics, DOI: 10.1080/00140139.2015.1122234

To link to this article: http://dx.doi.org/10.1080/00140139.2015.1122234

ERGONOMICS, 2016
http://dx.doi.org/10.1080/00140139.2015.1122234

Taylor & Francis
Taylor & Francis Group

The stress and workload of virtual reality training: the effects of presence, immersion and flow

S. J. Lackey[a], J. N. Salcedo[a], J.L. Szalma[b] and P.A. Hancock[a,b]

[a]Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA; [b]Department of Psychology, University of Central Florida, Orlando, FL, USA

ABSTRACT

The present investigation evaluated the effects of virtual reality (VR) training on the performance, perceived workload and stress response to a live training exercise in a sample of Soldiers. We also examined the relationship between the perceptions of that same VR as measured by engagement, immersion, presence, flow, perceived utility and ease of use with the performance, workload and stress reported on the live training task. To a degree, these latter relationships were moderated by task performance, as measured by binary (Go/No-Go) ratings. Participants who reported positive VR experiences also tended to experience lower stress and lower workload when performing the live version of the task. Thus, VR training regimens may be efficacious for mitigating the stress and workload associated with criterion tasks, thereby reducing the ultimate likelihood of real-world performance failure.

Practitioner Summary: VR provides opportunities for training in artificial worlds comprised of highly realistic features. Our virtual room clearing scenario facilitated the integration of Training and Readiness objectives and satisfied training doctrine obligations in a compelling engaging experience for both novice and experienced trainees.

ARTICLE HISTORY
Received 3 July 2015
Accepted 15 November 2015

KEYWORDS
Stress; workload; virtual reality; presence; immersion; flow

Introduction

Learning which is transferred from virtual realities (VR) appears to hold particular promise as a conduit through which training can be significantly enhanced (Arthur, et al., 1996; Kozak et al. 1993). We now see a variety of such capacities burgeoning in the realms of formal procedural training, as well as in the more expansive vistas of education in general. For example, Second Life (SL) has proved to be a highly influential development and has served as both a platform for, and an inspiration of the Military Open Simulator Enterprise Strategy (MOSES; Maxwell and McLennan 2012). This latter program represents an experimental simulation test bed for determining a set of best practices amongst all applicable military VR systems. Any such ideal system should improve on current training technology by offering realistic and persistent worlds in which events in that VR continue regardless of user connectivity. Such avatar-based VR's could be tailored to changing demands in training scenarios and would offer the capability to engage a great number of simultaneous users. All of these features are essential to the needs of future instructional programs and because of the opportunities in fluidity, adaptiveness and scalability involved in VR, it is perceived to be able to address the demands for ever more intricate and involved ranges of trained human behaviour (Maxwell and McLennan 2012).

For the MOSES project in particular to be meaningful, the associated VR must be technically feasible and give rise to improvement in the areas of its application. To enhance underlying technological progress, the Army Research Laboratory–Human Research and Engineering Directorate Simulation and Training Technology Center (ARL HRED STTC) MOSES community, and Intel Labs have focused on testing an alternative VR architecture involving a Distributed Scene Graphics. Such experiments have shown that the latter system possesses adequate management of computing load and bandwidth during distributed large user activity (with up to 60 simultaneous participants). It has thus generated data to assist in creating a predictive model concerning the performance effects on Soldiers of increased cognitive load (Maxwell, Geil, et al. 2014; Maxwell, Liu, et al. 2014). With respect to training applications, recent research has evaluated the effectiveness of such VR training by specifically comparing the performance of participants subjected to traditional live

CONTACT P.A. Hancock peter.hancock@ucf.edu

training with those who experience VR training in an Army room clearing task. These results demonstrated the efficacy of VR training for room clearing tasks in terms of performance and stress, giving promise for studying further military training tasks via VR (Lackey et al. 2014). Another training application for MOSES is its implementation at educational institutions. For instance, Tulane University has applied the MOSES Open Simulator platform as part of a proof of concept for VR in distance learning, with students finding great benefit in the interactive resources and tools therein (Maxwell and McLennan 2012).

The most widespread use of VR for education and training is through SL, which is a virtual 3D multi-participant online community that supports user-created environments, objects and avatars. Regions in its virtual world can be owned by users, who can share their creations and interact with other users. Companies including Crompco Corp., Walmart and Intel have used features of SL for internal group or individual employee training (Hof 2006). Organisations such as NASA and the US Army have also adopted SL for educational purposes (Allen and Demchak 2011). With respect to schools, most applications of SL are found in institutions of higher education, and over 100 universities (online or real-world) have established regions for academic activities (Baker and Brusco 2011). Academic courses using SL as an instructional tool include, but are not limited to, psychology, law, communications, humanities, history, nursing, architecture and writing (Atkinson 2008; Baker and Brusco 2011; Baker, Wentz, and Woods 2009; Calongne and Hiles 2007; Morgan 2013). Learning activities include simple virtual lectures, but a review by Inman, Wright, and Hartman (2010) showed that the educational tools also include role-play, group projects, student meetings, participating in simulations, participating in a community and more general social interaction.

In the light of this burgeoning form of technological augmentation to learning and training, the purpose for the present investigation was to evaluate the effectiveness of VR training relative to live training for an Army room clearing task using a sample of professional Soldiers as participants. In addition to objective performance, perceived workload and stress response, we were also interested in evaluating the effect of the VR on the perceived experience of the training. We thus measured self-reports of engagement, presence, immersion, flow state and the technology acceptance model (TAM: Venkatesh 2000). Measures of perceived usefulness and perceived ease of use of the VR during the training task were also elicited to evaluate the subjective experience alongside the performance output (Hancock 1996).

Experiment

Experimental participants

A total of 64 male US Army National Guard reserve Soldiers were recruited for this study. They were organised into 16, four Soldier Fire Teams. Participants were recruited through training personnel from the Army post facility, which also served as the experiment site. Each squad, consisting of two Fire Teams, was randomly assigned to one of two experimental conditions – VR training or Live training. The scope of the analyses in the current study focuses on the VR training condition. There were 32 participants in the VR condition (8 Four Soldier Fire Teams) aged 20–32 years ($M = 25.63$, SD = 3.14).

Experimental measures

Self-report data were collected using the empirically validated self-report measures which were collected in the order reported below.

Dundee stress state questionnaire

Pre- and post-training questionnaires for both VR and the Live training tasks were given in the form of the abbreviated version of the Dundee stress state questionnaire (DSSQ). The DSSQ helps measure changes in stress via three 0–32 point scales that comprise task engagement, distress and worry (see Matthews et al. 2013).

Cognitive workload

The NASA Task Load Index (NASA-TLX; Hart 2006; Hart and Staveland 1988) was used to assess participants' subjective workload experience during VR training. The NASA-TLX consists of six scales regarding workload (i.e. mental demand, physical demand, temporal demand, performance demand, effort, and frustration) rated individually on a 0–100 scale. The individual scales are aggregated for a Global Workload score.

Assessment of flow

Flow was measured with the flow state short scale (Jackson, Martin, and Eklund 2008). Participants were asked to rate (on a five-point rating scale) their experience of flow during the task across nine main aspects of flow: challenge and skill balance, action-awareness merging, clear goals, unambiguous feedback, concentration on the task at hand, sense of control, loss of self consciousness, transformation of time, and Autotelic experience. *Autotelic* refers to intrinsically satisfying and enjoyable experiences (Jackson and Marsh 1996).

Presence

Presence was measured using the questionnaire developed by Witmer and Singer (1998). Participants responded to 29 items on a seven-point rating scale ranging from 'Not Compelling' to 'Very Compelling'. The Presence Questionnaire is comprised of four scales, identified as Involvement and Control, Natural Interaction, Resolution, and Interface Quality.

TAM measures

The TAM measures of Perceived Usefulness and Perceived Ease of Use were measured using the scale derived by Davis (1989) and used more recently by Zhang, Li, and Sum (2006; see also Venkatesh 2000). These scales are each comprised of four items to which participants respond by rating the degree of agreement with the statements on a seven-point scale.

Engagement

Engagement in the VR was assessed using the scale developed by Charlton and Danforth (2005), which consists of seven statements regarding engagement in a virtual environment to which participants rate their agreement on a five-point scale.

Immersion

Using an immersion measure consisting of eight items selected from a scale developed by Jennett et al. (2008), participants rated the degree of immersion experienced in the VR on a five-point scale.

Experimental procedure

Each Soldier, who was recruited in the experiment, was notified that their involvement was voluntary, and explicit consent was elicited from all Soldier participants. Further, all Soldiers were allowed to discuss any relevant issues with project investigators, as well as review the research objectives, before choosing whether to participate or not. Those who subsequently did proceed could also decide whether they wished to be photographed and/or videotaped or not. If a Soldier chose not to participate, they were told that they may inform any investigator privately of their decision. In such cases, the investigator was instructed to tell the Soldier's unit supervisor, in brief terms, that the Soldier was unable to meet the research evaluation criteria and so was dismissed from further activity without prejudice.

Before the training phase, participants completed a questionnaire that captured their military training, military experience level, education and the frequency of video game usage. Participants were assigned at random to either the Live or the VR training conditions. The Live condition involved a traditional lecture and presentation which was supported by military doctrine (e.g. ARTEP 7-8-DRILL and FM 3-21.8), and included current room clearing knowledge. The latter information was reviewed and screened prior to the study by subject matter experts (SMEs). The material was presented by a Soldier on active duty who had expert knowledge of such room clearing tasks, including successful task execution and task evaluations in the field.

The room clearing training for the VR condition was preceded by general VR usage training to help acclimate the participants to the simulation itself. Similar to the Live condition, the VR room clearing training content included existing training doctrine (e.g. ARTEP 7-8-DRILL and FM 3-21.8) and room clearing content that was also vetted by SMEs. However, training materials were delivered through a computer simulation. Participants worked collaboratively within their Fire Teams for a maximum of five training trials, with each trial involving the completion of a room clearing scenario. Within their Fire Team, each participant was designated a certain role, and the Heads Up Display for each participant only showed the specific functions that were assigned to their current role.

After training sessions, a final assessment scenario was required for the participants from the Live and VR conditions. Participants were informed by the experimenter that the room clearing assessment matched the scenarios shown in training. Using a real-world room, participants within Fire Teams experienced a maximum of two room clearing scenarios for post-training performance assessment. After each scenario was completed, a binary performance grade (e.g. Go/No-Go) was provided by an instructor to each participant. When the final assessment was completed, each participant was offered a copy of their consent form. All participants in the VR condition were evaluated for symptoms of simulator sickness. If an instance of simulator sickness occurred, the participant was monitored at the research site, and dismissed only after their symptoms had dissipated.

Results

The relationships among the variables were evaluated via correlation and regression analyses. Correlations were computed to evaluate the direct relationship between each outcome variable (perceived workload and stress) and the predictor variables. Logistic regressions were computed for performance, and for perceived workload and stress. Hierarchical linear regressions were computed that included the predictor variable, performance outcome and the product vector of the predictor and performance. The predictors were scores on the measures of Engagement, Immersion, the four Presence scales, the TAM measures of Perceived Ease of Use and Perceived Usefulness, and

Table 1. Correlations of presence, engagement, immersion, ease of use, and usefulness scores with pre-task state.

	Pre task-engagement	Pre distress	Pre worry
Engagement	0.39*	0.08	0.30
Immersion	0.50**	0.07	0.53**
TAM scales			
Perceived ease of use	0.10	−0.48**	−0.45**
Perceived usefulness	0.42*	0.13	0.52**
Presence scales			
Involvement and control	0.39*	−0.20	0.20
Natural interaction	0.24	−0.20	0.31
Resolution	0.38*	−0.39*	−0.12
Interface quality	0.45**	−0.33	−0.18

*$p \leq 0.05$.
**$p \leq 0.01$.

Table 2. Correlations of flow state scale scores with pre-task state.

	Pre task-engagement	Pre distress	Pre worry
Challenge/skill balance	0.14	−0.36*	−0.43*
Action-awareness merging	−0.14	−0.40*	−0.43
Clear goals	−0.02	−0.42*	−0.29
Unambiguous feedback	0.18	−0.34	−0.24
Concentration on task at hand	0.36*	−0.19	−0.28
Sense of control	0.40*	−0.50**	−0.41*
Loss of self-consciousness	−0.006	−0.05	0.03
Transformation of time	0.18	0.06	0.27
Autotelic experience	0.49**	0.006	0.40*

*$p \leq 0.05$.
**$p \leq 0.01$.

the nine scales of the Flow measure. All regressions and correlations were computed using the residualised scores obtained after regression of each continuous measure on team membership. Residualised scores were used so that variance due to membership on a team would be removed prior to analyses.

Correlational analyses

DSSQ scales

The correlations between the pre-task DSSQ scales and the predictor variables are summarised in Tables 1 and 2. The corresponding partial correlations between post-task DSSQ scales and the predictor variables (with the variance due to the corresponding pre-task state removed) are summarised in Tables 3 and 4. Engagement, Immersion and Presence were most strongly linked to pre- and post-task engagement, although Engagement and the Presence scales of Involvement and Control and Interface Quality were each negatively related to post-task worry and post-task distress. Three of the four Presence scales were related to higher pre-task engagement, but only Involvement and Control was correlated with post-task engagement.

The TAM measures showed different patterns of relationship to stress. Ease of use was negatively correlated with pre- and post-task distress and worry, but perceived usefulness was correlated only with less pre-task worry and greater post-task worry. In contrast, ease of use was not significantly correlated with task engagement, but usefulness was positively correlated with both pre- and post-task scores on this measure. Among the flow scales, concentration was most strongly related to stress response, with substantial correlations across the three DSSQ scales. Note that Autotelic experience was related only to the motivation/energetic dimension of stress (i.e. task engagement) and (negatively) with the cognitive dimension (worry). Distress was related to the feedback, concentration and sense of control scales. In general, the correlations were in the expected direction. The opposite direction for the usefulness and pre-/post-task worry may have resulted from the stress of anticipation during the pre-task state.

NASA TLX scales

The correlations between the NASA TLX scales and the predictor variables are summarised in Tables 5 and 6. Across perceived workload dimensions, the TAM scale of perceived ease of use showed that the easier that the virtual training procedures were, the lower the perceived workload in using them. Presence scales were most strongly related to the stress dimension of the TLX, such that higher levels of presence reduced the frustration associated with using the technology. Among the flow scales challenge/skill balance, action/awareness merging, clear goals and sense of control were most strongly related to workload.

Table 3. Partial correlations of presence, engagement, immersion, ease of use, and usefulness scores with post-task DSSQ scores, controlling for pre-task state.

	Post task-engagement	Post distress	Post worry
Engagement	0.70***	−0.26	−0.53**
Immersion	0.44*	−0.03	−0.29
TAM scales			
Perceived ease of use	0.19	−0.64***	−0.49**
Perceived usefulness	0.57***	−0.10	−0.51**
Presence scales			
Involvement and control	0.48**	−0.42*	−0.38*
Natural interaction	0.10	0.01	−0.06
Resolution	0.01	−0.15	0.24
Interface quality	0.29	−0.51**	−0.36*

*$p \le 0.05$.
**$p \le 0.01$.
***$p \le 0.001$.

Table 4. Partial correlations of flow state scale scores with, post-task DSSQ scores, controlling for pre-task state.

	Post task-engagement	Post distress	Post worry
Challenge/skill balance	0.07	−0.52**	−0.40*
Action-awareness merging	−0.39*	−0.57	0.02
Clear goals	−0.12	−0.18	−0.16
Unambiguous feedback	0.12	−0.56***	−0.31
Concentration on task at hand	0.59***	−0.62***	−0.46**
Sense of control	0.16	−0.70***	−0.24
Loss of self-consciousness	0.13	−0.09	−0.20
Transformation of time	0.32	−0.03	−0.10
Autotelic experience	0.55***	−0.12	−0.35*

*$p \le 0.05$.
**$p \le 0.01$.
***$p \le 0.001$.

Table 5. Correlations of presence, engagement, immersion, ease of use, and usefulness scores with NASA TLX scores.

	Global WL	MD	PD	TD	PW	Effort	*F*
Engagement	−0.20	0.27	−0.36*	−0.11	−0.28	−0.21	−0.26
Immersion	−0.06	0.40*	−0.32	0.08	−0.19	−0.13	−0.30
TAM scales							
Perceived ease of use	−0.74***	−0.41*	−0.50**	−0.45**	−0.25	−0.66***	−0.66***
Perceived usefulness	0.02	0.51**	−0.29	0.09	−0.39*	0.16	−0.20
Presence scales							
Involvement and control	−0.39*	0.10	−0.33	−0.20	−0.44*	−0.24	−0.51**
Natural interaction	−0.08	0.28	0.01	0.01	−0.22	−0.11	−0.41*
Resolution	−0.38*	−0.14	−0.22	−0.24	−0.05	−0.35*	−0.57***
Interface quality	−0.18	0.10	−0.03	−0.29	0.001	−0.27	−0.32

Note: WL = Workload; MD = Mental demand; PD = Physical demand; D = Temporal demand; PW = Performance workload; F = Frustration.
*$p \le 0.05$.
**$p \le 0.01$.
***$p \le 0.001$.

In each case, higher scale scores were associated with lower perceived workload. The exception to this trend was observed for the Autotelic experience scale, which was positively correlated with mental demand.

Logistic regressions

Logistic regression analyses indicated that none of the measures predicted performance outcomes ($p > 0.18$ in each case) or experience in building clearing training. Regressions of performance and building clearing training experience on pre-task DSSQ measures were also computed. There were no statistically significant regressions for performance ($p > 0.47$ in each case). However, pre-task distress was associated with training experience, $\chi^2(1) = 5.08$, $p = 0.024$, −2 Log likelihood = 62.17, Nagelkerke $R^2 = 0.12$, $b = -0.18$ (SE = 0.08), Wald's test (df = 1) = 4.46, $p = 0.035$, Exp(b) = 0.84. Specifically, participants who experienced

Table 6. Correlations of flow state scale scores with NASA TLX scores.

	Global WL	MD	PD	TD	PW	Effort	*F*
Challenge/skill balance	−0.45*	−0.35*	−0.40*	−0.35*	−0.02	−0.48**	−0.17
Action-awareness merging	−0.49**	−0.30	−0.08	−0.64***	−0.10	−0.40*	−0.34
Clear goals	−0.47**	−0.32	−0.14	−0.55***	−0.10	−0.47**	−0.18
Unambiguous feedback	−0.32	−0.18	−0.21	−0.28	−0.06	−0.33	−0.20
Concentration on task at hand	−0.34	−0.11	−0.32	−0.23	−0.19	−0.29	−0.20
Sense of control	−0.55***	−0.23	−0.23	−0.52**	−0.18	−0.49**	−0.50**
Loss of self-consciousness	−0.18	−0.12	−0.004	−0.13	−0.02	−0.19	−0.23
Transformation of time	−0.05	0.18	−0.30	−0.08	−0.16	−0.002	0.06
Autotelic experience	0.19	0.45**	0.03	0.32	−0.31	0.20	−0.02

Note: WL = Workload; MD = Mental demand; PD = Physical demand; TD = Temporal demand; W = Performance workload; F = Frustration.
*$p \leq 0.05$.
**$p \leq 0.01$.
***$p \leq 0.001$.

greater pre-task Distress were more likely to have not had experience in building clearing training. The regressions of training experience on pre Task Engagement and Worry were not statistically significant ($p > 0.12$ in each case).

Linear regression analyses

The relationship between each predictor variable (i.e. scales measuring engagement, immersion, perceived ease of use, perceived usefulness, presence, and flow state) and outcome measures (i.e. post-task stress and perceived workload) were analysed via regression. In addition, the joint effects of predictor variables and performance outcome were also tested in separate hierarchical regressions for each predictor/outcome measure combination. Only the first performance evaluation was analysed because in the VR training condition all teams received 'Go' score for the second evaluation (i.e. there was no variability in performance at the second evaluation for participants in that training condition). Regressions to examine joint effects of predictor measures and building clearing training experience were not computed because of the 32 participants in the VR training condition only five reported no training experience.

For the regressions involving each post-task DSSQ scale, the respective pre-task score entered at step 1, step 2 included performance, the predictor variable comprised step 3 and the product vector of the categorical variable (performance) and the predictor variable was entered at step 4. Based on recommendations in Pedhazur (1997) regarding tests of significance for differences among regression coefficients (i.e. testing product vectors or 'interaction' effects), a more lenient alpha level (0.10) was adopted for the product vector entered at step 4.

Regressions of DSSQ scales on predictor variables and performance

There were no significant predictor by performance interactions for immersion ($p > 0.62$ in each case) or perceived ease of use ($p > 0.17$ in each case). Although there were statistically significant product vectors for the regression of post task engagement on presence-resolution and of post worry on the loss of self-consciousness scale, in each case separate regressions for each performance group were not significant ($p > 0.09$ in each case). Significant interactions likely occurred because of non-significant regression weights in opposite directions.

Engagement

A significant engagement by performance product vector was observed for post-Task Engagement (see Table 7). There were no statistically significant product vectors for the other DSSQ scales ($p > 0.22$ in each case). Separate hierarchical regressions of post Task Engagement on pre-task engagement (step 1) and the engagement measure (step 2) were computed for each performance group. In each case, the ΔR^2 refers to variance in post-Task Engagement accounted for by the engagement measure after the pre-task state was accounted for in the model. Statistically significant regressions were observed for the No Go performance group and for the Go performance group. The interaction resulted from a stronger positive relationship for the Go group relative to the No Go group. In both performance groups, participants who found the virtual environment more engaging also reported higher task engagement, but this relationship was stronger for those who performed well (see Figure 1).

TAM usefulness

A significant interaction term for perceived usefulness and performance was observed for Post-task Worry (see Table 8). There were no statistically significant product vectors for the other DSSQ scales ($p > 0.14$ in each case). Separate hierarchical regressions of post-task Worry on pre-task Worry (step 1) and perceived usefulness (step 2) were computed for each performance group. Statistically significant regressions were observed for both the No Go

Table 7. Regressions of DSSQ scales on predictor variables and performance.

Predictor	Regression	R^2	ΔF	ΔR^2	*b* (SE)	β
Engagement						
TE × Performance	$F(4,27) = 14.38^{***}$	0.68	$\Delta F(1,27) = 3.29^{\#}$	0.04		
No Go	$F(2,13) = 9.34^{**}$	0.59	$\Delta F(1,13) = 6.85^{*}$	0.22	0.40 (0.15)	0.49^{*}
Go	$F(2,13) = 19.38^{***}$	0.75	$\Delta F(1,13) = 25.08^{***}$	0.48	0.89 (0.18)	0.78^{***}
TAM usefulness						
W × Usefulness	$F(4,27) = 17.07^{***}$	0.72	$\Delta F(1,27) = 3.00^{\#}$	0.03		
No Go	$F(2,13) = 10.50^{**}$	0.62	$\Delta F(1,13) = 8.39^{*}$	0.25	−0.57 (0.20)	-0.62^{*}
Go	$F(2,13) = 29.39^{*****}$	0.82	$\Delta F(1,13) = 2.97$	0.04	0.19 (0.11)	−0.23
Presence: Resolution						
TE × Performance	$F(4,27) = 4.22^{**}$	0.38	$\Delta F(1,27) = 3.64^{\#}$	0.08		
Flow: Loss of self-consciousness						
W × Performance	$F(4,27) = 12.08^{***}$	0.64	$\Delta F(1,27) = 3.76^{\#}$	0.05		
Flow: Sense of control						
D × Performance	$F(4,27) = 14.17^{***}$	0.68	$\Delta F(1,27) = 4.96^{*}$	0.06		
No Go	$F(2,13) = 4.80^{*}$	0.42	$\Delta F(1,13) = 9.34^{**}$	0.41	−3.16 (1.03)	-0.66^{**}
Go	$F(2,13) = 44.10^{***}$	0.87	$\Delta F(1,13) = 22.70^{***}$	0.22	−5.09 (1.07)	-0.74^{***}

Note: TE = Post-task engagement; post-task *W* = Worry; *D* = Post-task distress; TAM = Technology acceptance model.
$^{\#}0.05 < p < 0.10$.
$^{*}p < 0.05$.
$^{**}p < 0.01$.
$^{***}p < 0.001$.

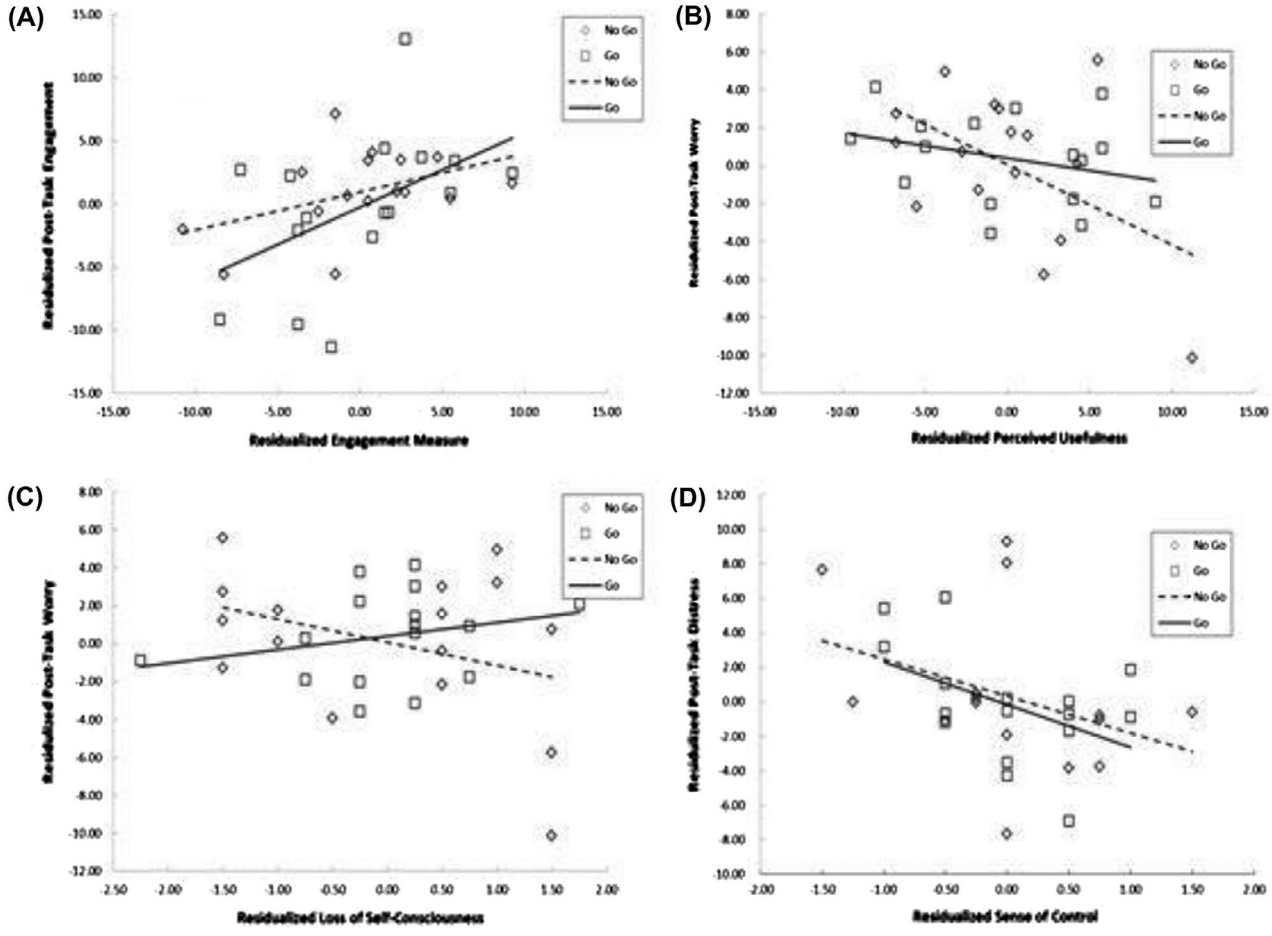

Figure 1. Residualised post-task DSSQ scores as a function of residualised flow scale scores for the Go and No Go performance groups.
Note: Scores are residualised after a regression of each measure on team membership.

Table 8. Regressions of NASA TLX scales on predictor variables and performance.

Predictor	Regression	R^2	ΔF	ΔR^2	b (SE)	β
Presence: natural interaction						
$F \times$ Performance	$F(3,28) = 3.49^{*}$	0.27	$\Delta F(1,28) = 4.04^{\#}$	0.10		
No Go	$F(1,14) = 11.96^{**}$	0.46			−3.95 (1.14)	−0.68**
Go	$p = 0.751$					
Flow: Challenge/Skill balance						
GWL × Performance	$F(3,28) = 3.72^{*}$	0.28	$\Delta F(1,28) = 3.32^{\#}$	0.08		
No Go	$p = 0.696$					
Go	$F(1,14) = 11.97^{**}$	0.46			−17.18 (4.96)	−0.68**
MD × Performance	$F(3,28) = 2.55^{\#}$	0.22	$\Delta F(1,28) = 3.28^{\#}$	0.09		
No Go	$p = 0.976$					
Go	$F(1,14) = 10.27^{**}$	0.42			−31.14 (9.72)	−0.65**
TD × Performance	$F(3,28) = 2.50^{\#}$	.21	$\Delta F(1,28) = 3.11^{\#}$	0.09		
No Go	$p = 0.995$					
Go	$F(1,14) = 10.20^{**}$	0.42			−22.10 (6.92)	−0.65**
PW × Performance	$F(3,28) = 1.14$	0.11	$\Delta F(1,28) = 3.40^{\#}$	0.11		
No Go	$F(1,14) = 4.90^{*}$	0.26			−19.80 (8.94)	−0.51*
Go	$p = 0.388$					
Eff × Performance	$F(3,28) = 15.74^{***}$	0.63	$\Delta F(1,28) = 30.12^{***}$	0.40		
No Go	$p = 0.236$					
Go	$F(1,14) = 91.36^{***}$	0.87			−41.43 (4.33)	−0.93***
$F \times$ Performance	$F(3,28) = 1.27$	0.12	$\Delta F(1,28) = 2.91^{\#}$	0.09		
No Go	$p = 0.555$					
Go	$F(1,14) = 4.75^{*}$	0.25			−13.33 (6.12)	−0.50*
Flow: Unambiguous feedback						
GWL × Performance	$F(3,28) = 2.42^{\#}$	0.21	$\Delta F(1,28) = 3.70^{\#}$	0.10		
No Go	$p = 0.930$					
Go	$F(1,14) = 7.16^{*}$	0.34			−14.37 (5.37)	−0.58*
MD × Performance	$F(3,28) = 3.66^{*}$	0.28	$\Delta F(1,28) = 9.68^{**}$	0.25		
No Go	$p = 0.276$					
Go	$F(1,14) = 12.59^{**}$	0.47			−32.18 (9.07)	−0.69**
PW × Performance	$F(3,28) = 1.83$	0.16	$\Delta F(1,28) = 5.38^{*}$	0.16		
No Go	$F(1,14) = 8.11^{*}$	0.37			−17.88 (6.28)	−0.61*
Go	$p = 0.259$					
Eff × Performance	$F(3,28) = 8.69^{***}$	0.48	$\Delta F(1,28) = 20.26^{***}$	0.38		
No Go	$p = 0.242$					
Go	$F(1,14) = 26.79^{***}$	0.66			−35.23 (6.80)	−0.81***
Flow: Sense of control						
MD × Performance	$F(3,28) = 1.84$	0.16	$\Delta F(1,28) = 3.77\#$	0.11		
No Go	$p = 0.834$					
Go	$F(1,14) = 6.65^{*}$	0.32			−25.42 (9.85)	−0.57*
Eff × Performance	$F(3,28) = 5.96^{**}$	0.39	$\Delta F(1,28) = 6.66^{*}$	0.14		
No Go	$p = 0.373$					
Go	$F(1,14) = 16.44^{**}$	0.54			−30.58 (7.54)	−0.74**

Note: GWL = Global workload; MD = Mental demand; TD = Temporal demand; PW = Performance workload; Eff = Effort; F = Frustration.
#$0.05 < p < 0.10$.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

performance group and for the Go performance group. The interaction resulted from a stronger negative relationship between perceived usefulness and post-task Worry for the No Go relative to the Go group. Thus, participants who perceived the VR training environment as more useful experienced less Worry, and this relationship was substantially stronger for those who did not perform well (see Figure 1(b)).

Flow

A significant interaction term for the Sense of Control scale and performance was observed for post-task distress (see Table 8). Separate analyses for each performance group indicated a significant regression for the No Go group and for the Go performance group. The significant product vector resulted from a stronger relationship between sense of

control and distress for the Go group. Participants across both groups who reported a higher Sense of control also reported lower levels of distress, but this relationship was stronger for those who performed well (see Figure 1(D)). There were no other significant product vectors involving flow measures and performance for the DSSQ scales ($p > 0.10$ in each case).

In sum, the relationships of stress to engagement, perceived usefulness and experience of Flow were moderated by performance outcome. Higher scores on the VR experience measures were associated with lower stress, but only for participants who performed well.

Regressions of NASA TLX scales on predictor variables and performance

A hierarchical regression was computed for each NASA TLX scale. In each case, step 1 consisted of performance, the predictor variable comprised step 2, and the product vector of the categorical variable (performance) and the predictor variable was entered at step 3. Based on recommendations in Pedhazur (1997) regarding tests of significance for differences among regression coefficients (i.e. testing product vectors), a more lenient alpha level (0.10) was adopted for the product vector entered at step 3. There were no significant predictor by performance interactions for engagement ($p > 0.45$ in each case), immersion ($p > 0.48$ in each case), perceived usefulness ($p > 0.24$ in each case) or perceived ease of use ($p > 0.11$ in each case).

Presence

A significant product vector involving the Natural Interaction scale and performance was observed for Frustration (see Table 8). Subsequent analyses indicated a significant regression for the No Go performance group but not for the Go performance group (for the latter condition $p = 0.751$). Higher Natural Interaction scores predicted lower frustration, but only for those who did not perform well. There were no other significant product vectors involving presence measures and performance for the TLX scales ($p > 0.24$ in each case).

Flow

Other than the effects summarised below there were no significant product vectors involving flow measures and performance for the TLX scales ($p > 0.10$ in each case).

Global workload

A statistically significant product vector was observed for the challenge/skill balance and unambiguous feedback scales (see Table 8). Subsequent analyses indicated that in each case, there was a significant regression for the Go performance group, but that the regressions for the No Go group were not significant ($p = 0.696$ and $p = 0.930$, respectively). Thus, for each of these flow scales, higher scores predicted lower global workload, but only for those participants who performed well.

Mental demand

Statistically significant product vectors were observed for the challenge/skill balance, unambiguous feedback, and sense of control scales (see Table 8 and Figure 2). Subsequent analyses indicated a significant regression in each case for the Go performance group, but the regressions for the No Go group were not statistically significant ($p = 0.976$, $p = 0.276$, and $p = 0.834$, respectively). Hence,

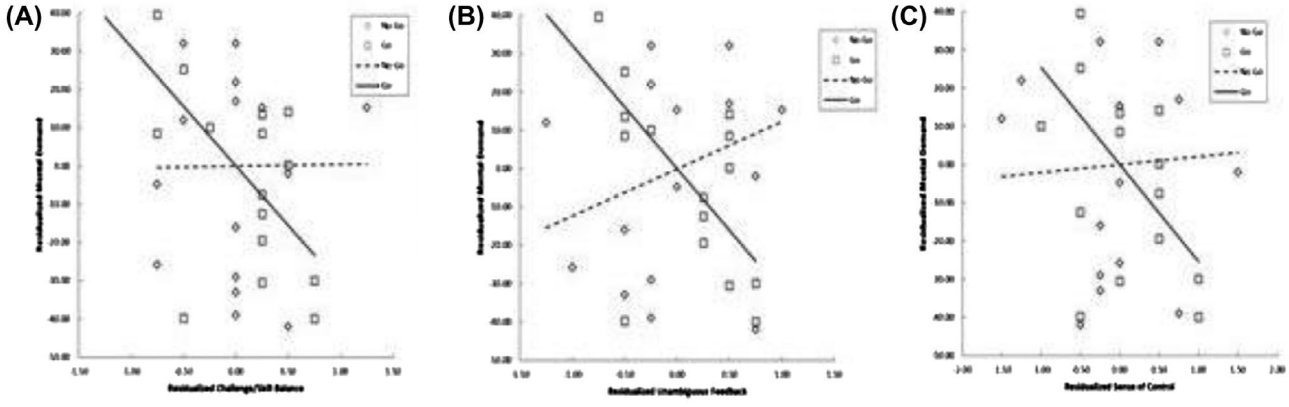

Figure 2. Residualised mental demand as a function of residualised challenge/skill balance (A), unambiguous feedback (B), and sense of control (C) flow scales.
Note: Scores are residualised after a regression of each measure on team membership.

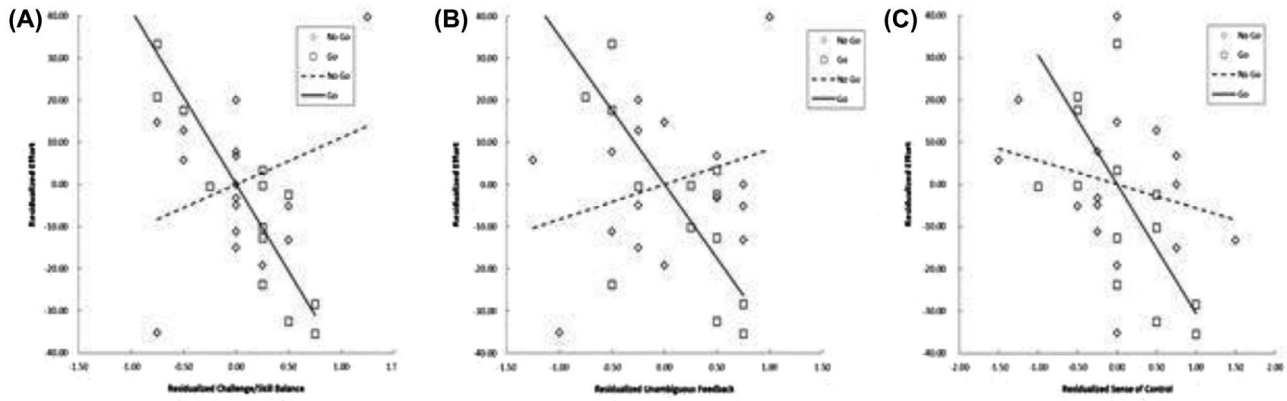

Figure 3. Residualised effort scores as a function of residualised challenge/skill balance (A), unambiguous feedback (B), and sense of control (C) flow scales.
Note: Scores are residualised after a regression of each measure on team membership.

for each flow scale higher scores predicted lower mental demand for those who performed well but not for those who performed poorly.

Temporal demand

A statistically significant product vector was observed for the Challenge/Skill Balance measure (see Table 8). Subsequent analyses indicated a significant regression for the Go performance group. The regression for the No Go group was not significant ($p = 0.995$). As in the case of the other TLX analyses, lower demand was associated with higher flow scores only for participants who performed well.

Performance workload

Statistically significant product vectors were observed for the challenge/skill balance and unambiguous feedback scales (see Table 8). Subsequent analyses indicated that in each case, there was a significant regression for the No Go performance group. The regressions for the Go group were not significant ($p = 0.388$ and $p = 0.259$, respectively). In contrast to mental demand, temporal demand, and global workload, for performance workload a significant negative relationship was observed only for those participants who performed poorly.

Effort

Statistically significant product vectors were observed for the challenge/skill balance, unambiguous feedback and sense of control scales. Subsequent analyses indicated that in each case, there was a statistically significant regression for the Go performance group (see Table 8 and Figure 3), but that the regressions for the No Go group were not significant ($p = 0.236$, $p = 0.242$, and $p = 0.373$, respectively). For each of the three flow scales greater flow scores were associated with lower perceived effort for those who performed well.

Frustration

A statistically significant product vector was observed for the challenge/skill balance measure (see Table 8). Subsequent analyses indicated a significant regression for the Go performance group, but the regression for the No Go group was not significant ($p = 0.555$). For those who performed well, the experience of a match between challenge and skill predicted lower frustration.

In sum, flow scales were related to lower global workload, although the patterns varied across the TLX subscales. Three components of flow emerged as predictors of perceived workload as a function of performance group: challenge/skill balance, unambiguous feedback and sense of control. These aspects of flow were associated with lower mental demand, temporal demand, effort and frustration, but only for participants who performed well. However, performance workload was lower as a function of Challenge/Skill Balance and Unambiguous Feedback for participants who did not perform well. Participants who performed well and also reported experiencing higher levels of challenge/skill balance, unambiguous feedback and sense of control in VR training reported lower perceived workload in the subsequent live training task.

Discussion

The primary goal for the present investigation was to evaluate the effects of VR training on the performance, perceived workload and stress response to a live training exercise. We also looked to determine how perceptions, as

measured by engagement, immersion, perceived usefulness and ease of use, presence, and flow, of the VR training task affected the performance, workload and stress of live task performance after training in VR. The VR experience measures did not predict performance, possibly because of the limited precision of the latter, i.e. being a dichotomous Go/No Go performance outcome. However, the stress and workload associated with the live training task was reduced by the experience of engagement, immersion, perceived usefulness and ease of use, and aspects of flow induced using VR for training. In several instances, these relationships were moderated by task performance. These prove to be examples of differing forms of performance–workload associations, dissociations and insensitivities (Hancock 1996; Hancock et al. 1995).

In general, the relationships of the predictor variables to workload and stress were in the hypothesised direction. Engagement was more strongly related to post-task stress and perceived workload, while presence and immersion scales were more strongly related to pre-task state (with the exception of the relationship of presence scales to frustration). TAM measures were correlated with both pre- and post-task stress and with cognitive workload, although the stress and workload dimensions were different for ease of use than for the usefulness scale. This pattern implies that experiences related to the usability of the VR technology primarily influenced the affective dimension of stress (distress), while perceptions of usefulness were related to the motivational/energetic aspect of stress (task engagement). The relationships of flow to stress varied as a function of scale. Challenge/skill balance and sense of control were associated with less distress but were unrelated to task engagement, while Autotelic experience was related to task engagement but not to distress. Concentration was related only to pre-task engagement but to all three DSSQ dimensions for the post-task state.

TAM measures

Perceived workload was more strongly related to usability than to perceptions of usefulness. Indeed, for the latter scale, higher scores were associated with greater mental demand as well as lower performance workload. However, these latter results may be due to the greater task engagement reported by those with higher perceived usefulness scores. That is, those who found the VR useful reported greater mental demand but also greater task engagement, possibly because they saw the value in engaging in the task. The positive correlation between mental demand and autotelic experience is also consistent with this interpretation. Perceptions of technological usefulness reduced worry, but more so for the participants who did not perform well. Perhaps the belief in the usefulness of the VR, combined with poor performance on the criterion task, induced less worry because the usefulness may have made the Soldiers feel that their poor performance would serve as useful feedback because the VR was useful.

Flow scales

Challenge/skill balance was the flow scale most strongly related to perceived workload, as higher scores were associated with lower global workload as well as lower ratings of task demand (mental, physical, temporal) and effort. With respect to stress, participants who reported a higher sense of control also reported lower distress, but this relationship was stronger for those who performed well. Performing well may thus facilitate both efficacy and level of stress associated with the task. In general, performing well allowed the experience of flow to be linked to lower workload, but performing poorly disrupted this relationship.

There were three dimensions of flow that seemed to drive workload: Skill/Balance, Feedback and Sense of Control. These relationships were only observed when performance was good, and for appraisals of task demand, effort and frustration. This is perhaps unsurprising, given that flow states are associated with high levels of performance efficiency. For these dimensions of workload, poor performance attenuated the flow–workload relationships. In addition, in each case greater flow scores were associated with lower perceived effort for those who performed well. This is consistent with a flow state being associated with more 'effortless' performance of an activity (cf., Hockey 1997 'effort without distress' control mode).

In contrast to the effects for the other workload scales, for perceptions of performance the negative relation of flow to workload occurred only for those who performed poorly. This may be due to performance workload reflecting appraisals of one's own response capacities, so those who performed poorly may have been more sensitive to the challenge of the task and the feedback they received regarding their performance. In this case, flow may have served to distract attention from thoughts regarding performance or to bias appraisals of performance, such that greater flow reduced performance workload even though actual performance was poor. Perhaps this dimension of flow serves actually to reduce the accuracy of perceptions of one's own performance. For those who performed well, the experience of a match between challenge and skill predicted lower frustration, as a flow theory perspective would predict. It is unfortunate that training experience could not be evaluated in these regressions because those with such experience and those who perform well might tend to have higher scores on flow scales. Future research should investigate this possibility.

Presence, immersion, and engagement

The presence and immersion scales were generally only weakly related to stress and workload. Immersion and three of the four presence scales were positively related to pre-task engagement, but only immersion and the Involvement and Control Presence Scales were related to post-task engagement. The Involvement and Control scale was also associated with lower post-task distress and worry, suggesting that this dimension of presence may have the greatest influence on the stress associated with performance after VR training. Interface Quality was associated with lower distress and worry, suggesting that positive perceptions of the quality of the technology used in training can reduce the stress of performance the live criterion task. By contrast, the natural interaction and resolution scales were unrelated to post-task stress, and immersion was related only to post-task engagement.

All four presence scales exhibited at least one significant correlation with perceived workload, although in all cases the relationships were to the TLX scales that assess perceptions of the self (i.e. performance workload, effort, and frustration) but not to appraisals of task demand (mental, physical, and temporal demand). Note that these latter ratings referred to the live criterion task. The Natural Interaction scale predicted lower frustration for participants who did not perform well. This factor of presence may thus attenuate the frustration experienced in the live criterion task for those who did not perform it well. It may be that in these cases the higher degree of presence directed attention towards the VR interaction and away from the self and thoughts regarding one's own performance, wherein those who experience a lower level of natural interaction may focus more on their poor performance and thus experience greater frustration. Engagement in the VR was related to greater post-task engagement, as one might expect, but this relationship was stronger for participants who performed well. Thus, the relationship of engagement in the VR to engagement in the live criterion task was enhanced for those who performed well on the latter.

Conclusions

The results of the present investigation indicated that self-report measures associated with how individuals experience a VR training task can predict the stress and workload imposed by a subsequent live transfer task but not the performance on that transfer task. Participants who reported positive experiences (e.g. usability, usefulness, flow) also tended to experience lower stress and workload when engaged in a live version of the task. Thus, VR training regimens may be efficacious for reducing the stress and workload associated with criterion tasks, thereby protecting mental resources and reducing the likelihood of subsequent, chronic performance failure (Hancock and Warm 1989; Hockey 1997). Future work in this domain should seek to deploy more sensitive objective performance measures for criterion tasks to be completed in operational environments as well as looking to examine the detailed temporal etiology of performance gains and decrements following VR as opposed to traditional forms of training.

Acknowledgment

The views and conclusions contained in this document are those of the authors and should not necessarily be interpreted as representing the official policies, either expressed or implied, of ARL HRED STTC or the US Government. The US Government is authorised to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the US Army Research Laboratory – Human Research Engineering Directorate Simulation and Training Technology Center (ARL HRED STTC), in collaboration with the Institute for Simulation and Training at the University of Central Florida via Contract# W911NF12R0011.

References

Allen, P. D., and C. C. Demchak. 2011. "Applied Virtual Environments: Applications of Virtual Environments to Government, Military, and Business Organizations." *Journal of Virtual World Research* 4 (2): 1–24.

Arthur, E. A., P. A. Hancock, and S. T. Chrysler. 1996. "Perception of Spatial Orientation in Real and Virtual Environments." *Ergonomics* 40 (1): 69–77.

Atkinson, T. 2008. "Second Life for Education: Inside Linden Lab." *Techtrends: Linking Research & Practice to Improve Learning* 52 (2): 18–21.

Baker, J. D., and J. M. Brusco. 2011. "Nursing Education Gets a Second Life." *AORN Journal* 94: 599–605.

Baker, S. C., R. K. Wentz, and M. M. Woods. 2009. "Using Virtual Worlds in Education: Second Life® as an Educational Tool." *Teaching of Psychology* 36 (1): 59–64.

Calongne, C., and J. Hiles. 2007. "Blended Realities: A Virtual Tour of Education in Second Life." Proceedings of the TCC Worldwide Online Conference, Honolulu, HI, 70–90.

Charlton, J., and I. Danforth. 2005. "Distinguishing Addiction and High Engagement in the Context of Online Game Playing." *Computers in Human Behavior* 23 (3): 1531–1548.

Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly* 13: 319–340.

Hancock, P. A. 1996. "Effects of Control Order, Augmented Feedback, Input Device and Practice on Tracking Performance and Perceived Workload." *Ergonomics* 39: 1146–1162.
Hancock, P. A., and J. S. Warm. 1989. "A Dynamic Model of Stress and Sustained Attention." *Human Factors* 31 (5): 519–537.
Hancock, P. A., G. Williams, S. Miyake, and C. M. Manning. 1995. "Influence of Task Demand Characteristics on Workload and Performance." *The International Journal of Aviation Psychology* 5 (1): 63–86.
Hart, S. G. 2006. "Nasa-Task Load Index (NASA-TLX); 20 Years Later." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50: 904–908.
Hart, S. G., and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*, edited by P. A. Hancock and N. Meshkati, 139–183. Amsterdam: North Holland Press.
Hockey, G. R. J. 1997. "Compensatory Control in the Regulation of Human Performance under Stress and High Workload: A Cognitive–Energetical Framework." *Biological Psychology* 45: 73–93.
Hof, R. D. 2006, April. *My Virtual Life*. Accessed from Bloomberg Business Week: http://www.businessweek.com/stories/2006-04-30/my-virtual-life
Inman, C., V. H. Wright, and J. A. Hartman. 2010. "Use of Second Life in K-12 and Higher Education: A Review of Research." *Journal of Interactive Online Learning* 9 (1): 44–63.
Jackson, S. A., and H. W. Marsh. 1996. "Development and Validation of a Scale to Measure Optimal Experience: The Flow State Scale." *Journal of Sport & Exercise Psychology* 18: 17–35.
Jackson, S. A., A. J. Martin, and R. C. Eklund. 2008. "Long and Short Measures of Flow: The Construct Validity of the FSS-2, DFS-2, and New Brief Counterparts." *Journal of Sport and Exercise Psychology* 30 (5): 561–587.
Jennett, C., A. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. 2008. "Measuring and Defining the Experience of Immersion in Games." *International Journal of Human-Computer Studies* 66 (9): 641–661.
Kozak, J. J., P. A. Hancock, E. Arthur, and S. Chrysler. 1993. "Transfer of Training from Virtual Reality." *Ergonomics* 36 (7): 777–784.
Lackey, S., J., Salcedo, G. Matthews, and D. Maxwell. 2014. "Virtual World Room Clearing: A Study in Training Effectiveness." Proceedings of the Interservice/Industry Training, Simulation, and Education Conference. Arlington, VA: NDIA.
Matthews, G., J. Szalma, A. R. Pabganiban, C. Neubauer, and J. S. Warm. 2013. "Profiling Task Stress with the Dundee Stress State Questionnaire." In *Psychology of Stress: New Research*, edited by L. Cavalcanti and S. Azevedo, 49–90. Hauppauge, NY: Nova Science.
Maxwell, D., J. Geil, W. Rivera, and H. Liu. 2014. "A Distributed Scene Graph Approach to Scaled Simulation-based Training Applications." Proceedings of the Inter-service/Industry Training, Simulation, and Education Conference (IITSEC). Arlington, VA: NDIA.
Maxwell, D., H. Liu, R. Adams, and D. Lake. 2014. "Make Large-scale Virtual Training a Reality." Proceedings of the MODSIM World Conference, Hampton, VA.
Maxwell, D., and K. McLennan. 2012. "Case study: Leveraging government and academic partnerships in MOSES (Military Open Simulator [Virtual World] Enterprise Strategy)." Proceedings of the World Conference on Educational Multimedia, Hypermedia, and Telecommunications (EdMedia), Denver, CO, 1604–1614.
Morgan, E. J. 2013. "Virtual Worlds: Integrating Second Life into the History Classroom." *History Teacher* 46 (4): 547–559.
Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction.
Venkatesh, V. 2000. "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model." *Information Systems Research* 11 (4): 342–365.
Witmer, B., and M. Singer. 1998. "Measuring Presence in Virtual Environments: A Presence Questionnaire." *Presence: Teleoperators and Virtual Environments* 7 (3): 225–240.
Zhang, P., N. Li, and H. Sun. 2006. "Affective Quality and Cognitive Absorption: Extending Technology Acceptance Research." Proceedings of the Hawaii International Conference on System Sciences, Kauai, HI, 207–217.