Embedding Secondary Tasks in Video Games to Measure Real-Time Cognitive Load: An Approach to Developing Adaptive Video Games

David Sharek, Psychology Department
Eric Wiebe, Department of Math, Science, and Technology Education
North Carolina State University, Raleigh, NC 27695-7650 USA

The goal of this study was to evaluate the use of an embedded secondary task in a video game as a measure of real-time cognitive load. Results show that performance in the secondary task was negatively correlated with the level of difficulty a player experienced, however secondary task performance and summative cognitive load self-report measures were not correlated. The ability to capture real-time data during gameplay indicates that secondary task performance could be used as a more sensitive measure of cognitive load compared with the summative cognitive load scores. Embedding secondary tasks in video games has the potential to be used to provide input for adaptive algorithms that sustain player engagement during gameplay.

INTRODUCTION

In order to create an adaptive gaming environment with the primary goal of promoting and sustaining engagement, it is essential to monitor aspects of engagement such as challenge, affect, and cognitive load during game play. In the past, adaptive systems have commonly been built around real-time performance measures. When a player is doing well, the adaptive engine might provide a more difficult game level; spawning more enemies, for example, or creating more complex puzzles to complete.

Though performance measures can provide snapshots of how well a player is doing during gameplay, they only provide an indirect measure of how challenged or cognitively loaded a player may feel. In order to sustain engagement, a player must feel positive affect and consistently optimally challenged (Csikszentmihalyi, 1990; Whitton, 2011), something performance measures may not be able to reveal. It follows that the measure of cognitive load must be concurrent with game play if it is going to be used as part of an adaptive algorithm. There are very few empirical research studies on measuring real-time video game engagement from a cognitive standpoint and consequently little is known about the efficacy of using possible mechanisms such as embedded game elements to measure cognitive load.

Adaptive Gameplay

When implemented well, adaptive gameplay algorithms can be used to sustain a player's engagement, and therefore increase their overall enjoyment and satisfaction in a game. Adaptive gameplay relies on computational algorithms to determine the appropriate level of difficulty a player should experience. This can occur either between levels, or in more complex instances, during gameplay. An example of adaptive game balancing begins with the system's ability to recognize when a player is either over-challenged or under-challenged. This is typically carried out by analyzing a player's performance. For example, if a player is unable to get past a particular point in a game, certain elements may be manipulated to decrease the level of difficulty, such as reducing the number of spawning enemies. Alternatively, if the game is comprised of discrete levels, such as in puzzle-based games, the next level presented could simply be an easier puzzle. It is important to note that adaptive gameplay should be managed carefully so as to prevent players from predicting how the adaptive algorithm will behave. If a player is aware of how the adaptive algorithm works, they could either "game" the system, or worse, become disengaged due to a lack of immersion once the reality of the world is spoiled (Schell, 2008). In adaptive systems, optimized challenge levels may lead to higher engagement when the game mechanics retain a degree of program control based on the user's real-time interactions with the system. (Gilleade & Dix, 2004).

It is likely that using an embedded secondary task during gameplay will provide strong opportunities for player engagement because it can seamlessly be integrated into the adaptive game engine without disrupting the player's focus on gameplay. Predetermined, static gameplay is unlikely to be as effective of an approach because it is too rigid and does not adjust to the unique needs of each player. Similarly, user-controlled gameplay may provide a more refined gaming experience but placing the burden of choice on the player requires that users are able to make choices that optimize their experience. For some games, these

choices may be designed into the gameplay to provide a more natural and seamless experience. However, for other games such as puzzle games which are more closely related to online training, asking the user to select a new degree of difficulty after each level may unnecessarily increase cognitive load, create a point of disengagement, and fail to optimize the training opportunities.

Measuring Cognitive Load

The ability to unobtrusively measure cognitive load during gameplay is necessary to provide a more holistic look at a player's experience while they are playing a video game. Cognitive load is commonly measured using post-hoc subjective measurement methods such as the Subjective Assessment Technique (SWAT) (Reid & Nygren, 1988), the Workload Profile (Tsang & Velazquez, 1996), and the NASA-TLX (Hart & Staveland, 1988). The NASA-TLX has been used extensively to measure overall workload, in addition to diagnosing individual affective and cognitive components of load (Byers, Bittner, & Hill, 1989; Hart & Staveland, 1988; Wiebe, Roberts, & Behrend, 2010). Though such subjective measurements of cognitive load are typically easy to implement, and provide straightforward data for analysis, they do have limitations. Some of these problems include an unclear relationship between perceived mental effort and actual cognitive load (Brunken, Plass, & Leutner, 2003). The fact that they must be delivered as a post-hoc assessment reduces sensitivity and prevents the accurate identification of the exact cause of the reported cognitive load (Brunken, Seufert, & Paas, 2010). This is particularly problematic for gaming environments where the level of load is likely to vary over the course of the experience.

In order to overcome some of these limitations, a secondary task methodology for measuring cognitive load has been adopted by some researchers as a relatively reliable and acceptable cognitive load measurement method (Brunken, Paas, & Moreno, 2010; Brunken, et al., 2003; Ogden, Levine, & Eisner, 1979).

Video game environments lend themselves well to including embedded secondary tasks. For example, many adventure games provide status bars for a game character's health. If this status bar is not monitored during a challenging fight, a player may not realize that they should heal their character, therefore resulting in losing the fight due to the death of their character. Another example is a type of vigilance task where players are required to monitor a secondary display such as a radar screen. If an event occurs in this secondary screen, the players are required to interface with the game in a specific way such as firing a weapon.

The time between an event occurring on the secondary task and the player's reaction can be used as a measure of load. For example, the longer it takes a player to react, the lower their performance on the secondary task indicating that they were experiencing higher levels of cognitive load from the primary task.

In this exploratory study, reaction time performance in an embedded secondary task during gameplay was compared with a summative self-report measure of cognitive load. It is expected that the real-time measure will provide more useful and actionable data for an adaptive algorithm compared to the post-hoc cognitive load measure. Due to the dynamic nature of playing a video game with multiple levels which create a myriad of potential reference points of perceived load, it is hypothesized that a player's secondary task mean reaction time will not correlate with their self-reported levels of cognitive load.

METHOD

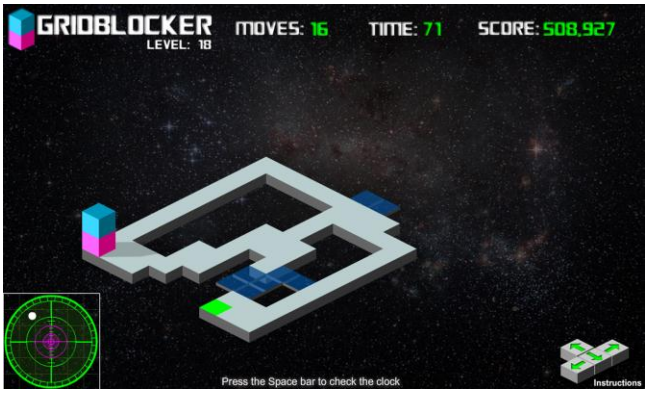
Participants

A total of 101 people were recruited through Amazon's Mechanical Turk (Amazon.com, 2009). Data from fourteen participants were removed due to a lack of interaction with the radar screen secondary task; they did not press the **F** key once during the game. To reduce the influence of practice effects, data from the first level everyone played was dropped. Of the remaining 87 participants, 47% were female and 53% were male (mean age = 32.99, *SD* = 10.43). Ninety-seven percent were from the United States. Thirty percent indicated that they had previous experience with a similar type of puzzle game.

Apparatus

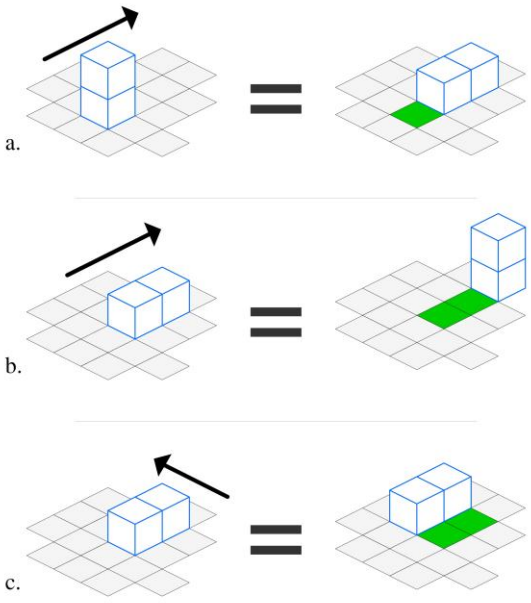
The experiment was hosted on www.gridblocker.com, a website developed with the sole purpose to create a unified gaming experience for the GridBlocker game. Backend data collection was managed via a PHP-MySQL configuration. The only requirements for accessing the online game was an Internet connection, a computer monitor resolution that could support at least 1000x600, and an installed version of the Adobe Flash player v.10 or greater.

Figure 1. Screenshot of game play in the video game, GridBlocker (Sharek, 2011b)



As can be seen in **Figure 1**, participants played the online Flash-based strategy game called GridBlocker (Sharek, 2011b). GridBlocker was specifically developed to be an experimental tool to measure various psychological constructs such as Flow, engagement, and cognitive load.

Figure 2. Moving the Block Around the Game Board in the GridBlocker Game (Sharek, 2011b)



The goal of the isometric tile-based game is to move a rectangular block, made up of two differently colored cubes, towards a goal point so that it is standing up on top of the goal. In more difficult levels, the goal will only accept the end of the block that is of the same color as the goal. As **Figure 2** shows, the movement of the block depends on its starting position on each tile. When the block is standing up, any horizontal or vertical movement will cause it to fall down. Notice in Move **a.** how the block no longer remains over the starting tile position (shown in green).

When the block is laying down, it can be moved to a standing position if the movement is along the longer axis of the block. A combination of Move **a.** and Move **b.** will move the block three squares along an axis

If the block is laying down and moved along the block's shorter axis, it will continue to remain laying down and only move one square at a time. Combinations of Move **a.** and Move **c**. can be learned to place the block in varying locations.

Design

The exploratory nature of the design consisted of a single linear increase in difficulty condition. Game levels were designed based on prior research studies (Sharek, 2010). Players began playing the game at a low difficulty level and each subsequent level became incrementally more difficult. This condition was based on findings from previous studies (Sharek, 2010; Sharek & Wiebe, 2011) that showed that Flow theory could be used to create an experience where a player's skill and the game's difficulty are managed and appropriately matched. This skill-difficulty match should allow the player to gain the skills necessary to overcome the game's increasingly difficult levels.

Performance Dependent Variables. The number of times a player moved the block over and beyond the calculated minimum number of times required to solve the puzzle (**Over Moves**), the number of times the block's direction changed (**Directions**), and the total length of time a player spent playing each level (**Time**) were analyzed as performance measures across levels. Including the aforementioned performance measures, a player's score at the end of the game (**Score**) and the number of levels a player completed (**levels**) was analyzed as individual performance measures across participants.

Cognitive Load Dependent Variables. A real-time measure of cognitive load was captured via response time in the secondary task radar screen (**RT**). A post-hoc summative measure of cognitive load was captured using the NASA Task Load indeX (**NASA-TLX**) (Hart & Staveland, 1988).

Procedure

The experiment was conducted over an Internet connection. Once participants provided consent for the terms of the task, they watched a brief tutorial that explained how to play the game, including how to interact with the secondary task radar screen. The experiment began after the tutorial was completed.

After playing the game, participants were asked to complete a demographics questionnaire and a computerized version of the NASA-TLX (Sharek, 2011a). When the questionnaire was completed, a final page was displayed that debriefed and thanked the participants. On this page, participants were also given an experimental completion code which they used to paste into their Mechanical Turk worker page to indicate that they have completed the experiment where they were then awarded payment in the amount of one dollar.

RESULTS

Performance and self-report measures are presented in this section.

A manipulation check was conducted to determine if the design condition accurately reflected the intended linear increase in difficulty. Pearson correlations revealed that level **difficulty** was positively and significantly correlated with: **Over Moves** ($r(43) = .46$, $p<.01$), **Directions** ($r(43) = .53$, $p<.001$), and **Time** ($r(43) = .49$, $p=.001$).

In a level-by-level analysis, a significant and positive Pearson correlation was found between response time (**RT**) in the secondary task radar screen and game **difficulty** level, $r(43) = .54$, $p<.001$. As difficulty increased, player's performance (higher response time = lower performance) in the secondary task decreased.

Summative player data were also analyzed using Pearson correlations. **NASA-TLX** scores were significantly and negatively correlated with both the maximum **difficulty** level a player was able to complete ($r(85) = -.32$, $p<.01$) and their **Score** ($r(85) = -.34$, $p=.001$). **NASA-TLX** scores were not correlated with mean participant secondary task response times (**RT**).

DISCUSSION

Data from the manipulation check provided evidence that the video game difficulty levels were designed as expected. The more difficult a level was, the more moves a player made over and beyond the solution, indicating that finding the optimal solution to levels became less clear as difficulty increased. As would be assumed, the more difficult a level was, the more directional changes a player made with the block. Finally, as difficulty increased, so did the amount of time it took a player to solve the level.

As expected, level-by-level secondary task response time was positively correlated with difficulty level. In other words, the more difficult a level was, the more time it took for a player to perceive an asteroid on the radar screen. This indicates that as the primary task (moving the block towards to goal) became more difficult in successive levels, it demanded greater amounts of cognitive effort compared to less difficult levels. This is a promising result because it shows that the secondary task can potentially be used to determine whether a level is cognitively underloading or overloading a player.

The NASA-TLX scores were negatively correlated with the maximum difficulty level a player was able to reach. On the surface, this seems like a reasonable result. Since game levels were served in a linear order of difficulty, better players would reach higher levels and it is likely that they also perceived lower overall cognitive load. The NASA-TLX scores were also negatively correlated with overall game scores. Again, better players would both receive higher scores and also may perceive lower load. However, the summative nature of the NASA-TLX removes the sensitivity required to determine exactly what aspects of the game influenced the rating scores. For example, a less likely alternate hypothesis for the above correlations is simply a function of a recency effect. The player's experience from the last game level they played or their final score may have influenced the self-report ratings, rather than their overall experience. The lack of a correlation between a player's NASA-TLX scores and their mean response times from the secondary task highlight this gulf between post-hoc and real-time data.

There are two major advantages to pursuing the development of real-time cognitive load data collection in video games. The first advantage is that data collected during gameplay could be analyzed during playtesting to help developers identify parts of a game that players find to be too easy or too difficult. This actionable data could be used to quickly influence game development decisions as part of an iterative design methodology such as Microsoft's Rapid Iterative Testing and Evaluation (RITE) method.

The second advantage to collecting real-time cognitive load data lies in the use of the data as part of an adaptive algorithm. Discreet levels could be served to

the players based on an algorithm using the real-time cognitive load data. For example, if a player is performing poorly in the secondary task (as indicated by higher response times), the algorithm could determine that the next level they play should be slightly less difficult. Incorporating performance data would provide an even clearer picture of the player's potential level of engagement. For example, if a player is performing poorly in the secondary task, but well in the primary task, the algorithm may determine that the level is maxing out the player's cognitive capacity. In this case, the next level served may need to be equally difficult, despite the poor performance in the secondary task.

The data show that the use of a secondary task as a real-time measure of cognitive load during video gameplay is promising. Future studies should be conducted to determine the efficacy of data from the secondary task as input into an adaptive algorithm. Of particular interest is exploring the notion of "achievable difficulties" from Flow Theory (Csikszentmihalyi, 1990; Sharek & Wiebe, 2011). If a user's expertise increases at the same rate as game level difficulty, one would expect no correlation between game difficulty level and secondary task reaction time, yet such a relationship was seen in this study. It may be that there would be a point at which this no longer functions as a linear relationship. An adaptive algorithm based on real-time measures would need to be able to detect these outer boundaries of achievable difficulties.

REFERENCES

Amazon.com. (2009). Mechanical Turk. Retrieved May 2, 2009, from https://www.mturk.com/mturk/welcome

Brunken, R., Paas, F., & Moreno, R. (2010). Current Issues and Open Questions in Cognitive Load Research. In J. Plass, R. Moreno & R. Brunken (Eds.), *Cognitive Load Theory* (pp. 253-272). New York: Cambridge.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. [Article]. *Educational Psychologist, 38*(1), 53-61.

Brunken, R., Seufert, T., & Paas, F. (2010). Measuring Cognitive Load. In J. Plass, R. Moreno & R. Brunken (Eds.), *Cognitive Load Theory* (pp. 181-202). New York: Cambridge.

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw Task Load Index (TLX) correlations: Are paired comparisons necessary? In A. Mital (Ed.), *Advances in industrial ergonomics and safety*. London: Taylor & Francis.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper and Row.

Gilleade, K., & Dix, A. (2004). *Using Frustration in the Design of Adaptive Videogames.* Paper presented at the Proceedings of Advances in Computer Entertainment Technology, Singapore.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland.

Ogden, G., Levine, J., & Eisner, E. (1979). Measurement of Workload by Secondary Tasks. *Human Factors, 21*, 529-548.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185-218). Amsterdam: Elsevier.

Schell, J. (2008). *The Art of Game Design: A Book of Lenses*. Burlington, MA: Morgan Kaufmann.

Sharek, D. (2010). *The Influence of Flow in the Measure of Engagement.* Unpublished Master's Thesis, NCSU, Raleigh, NC.

Sharek, D. (2011a). *Developing a Usable Online NASA-TLX Measurement Tool*. Paper presented at the Human Factors and Ergonomics Society 55th Annual Meeting.

Sharek, D. (2011b). GridBlocker (Version 1.0) [Computer Game]. Raleigh, NC.

Sharek, D., & Wiebe, E. (2011). *Using Flow Theory to Design Video Games as Experimental Stimuli*. Paper presented at the Human Factors and Ergonomics Society 55th Annual Meeting.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics, 39*(3), 358 - 381.

Whitton, N. (2011). Game Engagement Theory and Adult Learning. *Simulation & Gaming, 42*(5), 596-609.

Wiebe, E. N., Roberts, E., & Behrend, T. S. (2010). An Examination of Two Mental Workload Measurement Approaches to Understanding Multimedia Learning. *Computers in Human Behavior, 26*, 474-481.