Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

EXPLORING GAZE BEHAVIOR TO ASSESS PERFORMANCE IN DIGITAL GAME-BASED LEARNING SYSTEMS

Brian An
Inki Kim
Erfan Pakdamanian
Donald E. Brown

Department of Systems and
Information Engineering
University of Virginia
Charlottesville, VA 22904, USA

ABSTRACT

The recent growth of sophisticated digital gaming technologies has spawned an $8.1B industry around using these games for pedagogical purposes. Though Digital Game-Based Learning Systems have been adopted by industries ranging from military to medical applications, these systems continue to rely on traditional measures of explicit interactions to gauge player performance which can be subject to guessing and other factors unrelated to actual performance. This study presents a novel implicit eye-tracking based metric for digital game-based learning environments. The proposed metric introduces a weighted eye-tracking measure of traditional in-game scoring to consider the mental schema of a player's decision making. In order to validate the efficacy of this metric, we conducted an experiment with 25 participants playing a game designed to evaluate Chinese cultural competency and communication. This experiment showed strong correlation between the novel eye-tracking performance metric and traditional measures of in-game performance.

1 INTRODUCTION

The recent and rapid growth of the computer gaming industry has also served as a catalyst for the growth of an $8.1B industry around computer game-based pedagogical tools also known as digital game-based learning (DGBL) systems (Adkins 2017). DGBLs have gained popularity across a wide range of applications for a number of reasons but primarily due to their repeatability and cost-effective scalability. Despite becoming a critical component to the educational economy, there has yet to be discovered an effective and cross-domain method to measure performance within DGBLs.

Initial research in this area focused on the potential learning benefits of commercial off-the-shelf entertainment games, but recently many industries and researchers have transitioned to the development and evaluation of games specifically designed for learning (Subrahmanyam and Greenfield 1994; Connolly et al. 2012). A number of studies have aggregated and objectively evaluated games ability to meet specific learning objectives (Connolly et al. 2012). Common to many of these evaluation methods is the use of explicit user-response scores which may be susceptible to guessing, thus not accurately reflecting a player's knowledge and performance. Eye-gaze technology used in tandem with DGBLs provides a novel path to measure player performance.

Eye-tracking has been successfully used in activity recognition, affect detection, engagement, cognitive load assessment, and general learning (Bulling et al. 2011; Courtemanche et al. 2011; Jaques et al. 2014; Wang et al. 2014). Modern eye-tracking technologies allow sophisticated passive process tracking with minimal cognitive load of more traditional process tracking methods such as think-aloud protocols (Lohse

and Johnson 1996). Presupposing the validity of the Eye-Mind assumption (Just and Carpenter 1980), the high frequency spatial and temporal information available through eye-tracking present an avenue by which one can infer both conscious and subconscious cognitive processes that would be too ephemeral for active process tracking methods. The general Eye-Mind assumption asserts that the object of fixation is the focus of attention and cognitive processing. Widespread research has supported the Eye-Mind assumption, though a number of studies have also presented situations where it was shown to be questionable or insufficient (Kiili et al. 2014) especially in situations when the visual scene is unrelated to the intended cognitive activity.

This study contributes to the growing body of research investigating the eye-tracking behaviors as they relate to performance within DGBLs. Specifically, we aim to determine whether a player's eye-tracking behavior correlates to their performance in order to substantiate a novel eye-tracking based method for evaluating player performance. During this study, we evaluate the eye-tracking data of players in a DGBL designed to evaluate participants in Chinese Cross Cultural Competence.

2 BACKGROUND

Despite the advances in the operationalization of DGBLs, research scrutinizing the methods of measuring user performance has not demonstrated the same level of growth. A majority of DGBLs leverage post-game examinations of learning objectives in the form of pre/post testing (Bellotti et al. 2013). The actual implementation of the pre/post tests span a number of methods to include evaluating player's knowledge through questionnaires/surveys (Fishwick et al. 2010; Coffey et al. 2017). A common derivation of this method is the Situational Judgment Tests (SJT) designed to use the learned knowledge in practical applications (Lane et al. 2013). These methods have been deployed on a wide scale primarily for their ease of implementation, though they suffer from several notable shortcomings. Specifically, they fail to consider whether the pre-test influenced the post-test results as seen in the case of the Second China game experiment where the pre/post tests were the same test with the questions reordered (Coffey et al. 2017).

Psycho-physiological data as a means of assessing learning performance within DGBLs is a rapidly growing research area. Specifically, the use of eye-tracking has gained significant traction due to the advances in both eye-tracking technology as well as DGBL development tools. Companies such as Tobii, Pupil Labs, and SensoMotoric Instruments have significantly reduced the cost and increased the resolution of eye-tracking devices while also offering a variety of collection devices for diverse collection situations. Additionally, the proliferation of popular gaming stacks to include Unity3d and Unreal Engine have made it possible for smaller research efforts to develop high quality scalable game environments for minimal costs.

With this proliferation, various applications of eye-tracking in DGBLs have been explored to determine game effectiveness. One specific research area targets the measurement of player engagement and flow through the interpretation of affect-based metrics (Renshaw et al. 2009; Jaques et al. 2014; D'Mello et al. 2012). A limitation, however, of affect-based measures such as pupil dilation is that they are generally one dimensional and exhibit a variable lag therefore providing limited insight into complex cognitive processes associated with learning (Wierda et al. 2012).

In order to better investigate user performance, researchers have used fixation location and gaze paths to examine performance in DGBLs. Specifically, researchers have conducted experiments to identify differences in visual behavior based on skill level. Kickmeier-Rust et al. found significantly different gaze path and interaction strategies between high/low performance levels of teenagers playing a game designed to teach European geography (Kickmeier-Rust et al. 2011). In a study examining adolescent behaviors in attention enhancement therapy, Pascual et al. found that higher performance children exhibited lower fixation densities. Additionally, they were able to use ensemble machine learning methods on a combination of the gaze data and user interaction features to create a prediction model (Frutos-Pascual and Garcia-Zapirain 2015).

More recently, researchers have begun to investigate the use of eye-tracking to understand the process and rate of learning through gaze metrics (Wu et al. 2014; Józsa and Hámornik 2011). Józsa et al. studied the possibility of assessing a learning curve through trends in the fixation duration and total visit duration between successive stimuli. The results, however, were mixed due to the variability of the stimuli (Józsa and Hámornik 2011).

The overall purpose of this research effort is to introduce and evaluate the efficacy of a novel eye-tracking measure of performance in social-interaction DGBL systems. In doing so, we contribute to the rapidly growing need for effective methods of measuring performance in DGBL systems.

3 EYE-TRACKING MEASURE OF PERFORMANCE (ETMP)

Social Interaction-based serious games often have players interact with computer-driven avatars through finite dialogue trees. When a conversation commences, players are expected to sequentially select one of several dialogue options in order to progress through a conversation. In some cases, the dialogue options are either appropriate or inappropriate, but often, dialogue options have varying degrees of appropriateness scores which can have various impacts on the rapport that players can build with the non-player characters. In the latter case, there is an opportunity to leverage the range of appropriateness with visual attention in order to produce a weighted score of appropriateness as proposed below:

$$Player_{Score} = \sum_{i=1}^{x} \sum_{j=1}^{y} s_{ij} f_{ij}$$

where s_{ij} is the score of response option j in question i, f_{ij} is the fixation proportion on response option j relative to the other dialogue options in question i, y is the number of dialogue options in a conversation instance, and x is the total number of conversation instances in the conversation. By using fixation proportions instead of absolute fixation durations, the ETMP minimizes the impact of variable reading speeds of the participants. Research has shown a gaze bias in which people fixate on options with higher ratings of subjective preference (Glaholt and Reingold 2011; Shimojo et al. 2003). Assuming the validity of the previously mentioned Eye-Mind Theory (Just and Carpenter 1980) and gaze bias findings, this proposed formulation would conceptually evaluate both the appropriateness of a response as well as consider the mental schema the player uses to determine his or her response. It is important to note that this methodology only captures the overt movements of visual attention as measured through eye-tracking and makes no assumptions of the covert movements of attention that would not be observable through eye movements (Duchowski 2002). The gaze-based weighting of performance in this formulation results in higher scores for those players that spend more visual attention on the appropriate answers and lower scores for participants that dwell on less appropriate answers.

4 RESEARCH QUESTIONS

In order to investigate the feasibility and viability of a gaze-based measure of performance, this study examines how users distribute their visual attention across multiple dialogue options to progress conversations with avatars within a serious game designed to evaluate Cross Cultural Competence in Chinese Culture. The specific research questions and related hypotheses are as follows:

RQ-1) How do users distribute their visual attention across dialogue options in social interaction serious games? Do they focus their attention on preferred options and less so on rejected options?

- H_{1a}: The game participants will spend significantly more time fixating on selected answer choices than non-selected choices
- H_{1b}: The higher-scoring participants will spend significantly more time fixating on culturally appropriate responses

- H_{1c}: The advanced domain knowledge or metacognitive-ability participants will spend significantly more time fixating on higher-score responses

The intent of RQ-1 is to better understand the visual behavior of participants in a serious game designed to improve Cross Cultural Competence. Specifically, we investigate whether the visual behavior in this game is comparable to that of other research findings in both visual behavior in multiple choice environments and other DGBL systems.

RQ-2) Does the Eye-Tracking Measure of Performance correlate to other measures of performance? Does higher domain knowledge, gaze bias, or higher traditional game scoring correlate to higher scores in the ETMP?

- H_{2a}: The game participants with advanced domain knowledge or metacognitive skill will score significantly higher in the ETMP than those participants with lower domain knowledge or less skilled.
- H_{2b}: The Participants who scored higher in the traditional game scoring will also score significantly higher in the ETMP.
- H_{2c}: The participants who exhibited a gaze bias on their selected answer will score significantly higher in the ETMP.
- H_{2d}: The participants who exhibited a gaze bias on appropriate answers will score significantly higher in the ETMP.

The intent of RQ-2 is to provide initial validation to the ETMP through correlation to other indicators of performance whether they be in-game measurements or independent measures of Cross-Cultural Competence (3C). 3C continues to be an active area of research without a clear consensus on a 3C theoretical construct (Chiu et al. 2013). Though we discovered numerous 3C constructs and inventories in our review of the literature, we chose the Cultural Intelligence (CQ) inventory to assess an individual's 3C given promising results in recent independent validation efforts (Van Dyne et al. 2008; Matsumoto and Hwang 2013).

5 QUASI-EXPERIMENTAL DESIGN

In an effort to operationalize and validate the ETMP, we conducted an experiment with a DGBL system. The following section describes the quasi-experimental design, setup and analysis approach of the experiment (Cook et al. 2002).

5.1 Participants and Design

25 participants (40% male; M_{age}=20.67; SD_{age}=1.80) were recruited from the University of Virginia to participate in the experiment. All of the recruited participants were students in the Chinese Language Department. Each student was administered the Cultural Intelligence (CQ) Inventory to gauge self-reported 3C. 6 of the 25 participated in the study exceeded our 25% eye-tracking data loss threshold and thus had to be excluded from the analysis of the results resulting in a sample size of 19.

5.2 Materials and Measures

The stimulus in this experiment is a custom developed first-person serious game placed in a Chinese university called the Chinese Cultural Dimension Training System (CCDTS) (Bayer et al. 2017). In the game, the user is expected to collaborate with familiar and unfamiliar Chinese students and professors while working on a group project. As the story continues, users engage in various conversations concerning politics, cultural habits, and stereotypes concluding with the player coordinating meeting times with a fellow student and visiting office hours to clarify questions with the professor. The game begins with an initial in-game training segment which introduces the storyline, allows the user to become accustomed to the game controls, and provides specific instruction to Chinese culture. It is then followed by two evaluation

scenes called *Library* and *Office Hours*. In order to adapt the CCDTS to collect eye-tracking data, the bounding areas of each dialogue response had to be enlarged and had preset as Areas-of-Interest within the Tobii Pro Studio eye-tracking software.

5.2.1 Scene Descriptions

In the *Library* scene, the player must engage with a fellow student about the group project details and coordinate meeting times. In the *Office Hours* scene, the player engages with a professor during Office Hours to get formal feedback and recommendations about their group project.

5.2.2 In-Game Scoring

During the development of the game, it was necessary to score and validate the cultural appropriateness of specific dialogue responses by means of interrater reliability. For this purpose Chinese cultural experts were recruited from the Chinese language department. The raters were asked to rate each response on a likert scale of 1 to 5 with 1 being the least appropriate and 5 being the most appropriate. The ratings determined the explicit scores associated with each response. This score was used to measure the player's performance as this experiment was not designed for a pre-test to target learning gains. The CQ inventory administered pre-stimulus, however, allowed us to investigate any pre-game cultural competency. A more detailed discussion of the scoring methodology is described by Bayer et al. (2017).

Given this rating system, a response with a score of either 4 or 5 was considered to be an appropriate score. Additionally, if an interaction did not have a response scored as a 4 or 5, only the highest scored response in that interaction was considered as a 'appropriate score' in determining the total fixation durations.

5.3 Apparatus

The stimulus was presented on a 33-in screen with a pixel resolution of 2560 by 1440. The CCDTS was played through an executable Unity file and eye-tracking data was recorded using Tobii Pro Studio V1.1. During testing, participants were seated approximately 65 cm away from the screen. The vertical placement of the screen was adjusted such that the center of the screen was at eye level of each participant. Eye-tracking data was collected using the Tobii X3 120, a screen-mounted collection system. The system was calibrated for each participant using the Tobii Pro Studio animated 8 point calibration system. Calibration accuracy was recorded to be within 0.6°of visual angle for both axes of all participants.

5.4 Procedure

Students were individually tested. Initially they were administered a demographic survey followed by the CQ Inventory. The CQ Inventory was used as a means to assess a players pre-game cross-cultural competence as a possible predictor of the ETMP. They were then calibrated to the eye-tracking apparatus within the tolerance of the eye-tracker and the CCDTS was initiated. The first scene in the game was a basic tutorial to teach participants how to navigate the interface and play the game. Once the tutorial was completed, the participant immediately entered the real game environment in which they interacted with non-player characters using the multiple choice style dialogue system common to modern video games. This test flow is depicted in Fig 1. Following the completion of the game, a short after-action interview was conducted with the participant to attain any feedback about the game experience. The information collected from these interviews are not reported in this publication.

5.5 Data Collection & Analysis

Prior to analyzing the data, the fixation durations on specific areas of interest (AOI) had to be processed from the collected gaze data. Since each interaction presented a series of responses for the participant to evaluate, the container around each of the responses was considered an AOI. All fixations within the

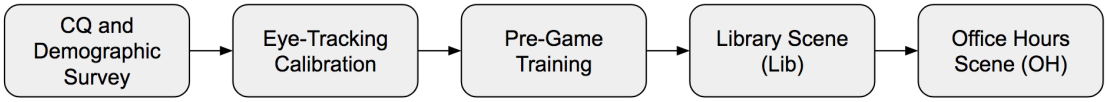

Figure 1: Experimental Design.

container were aggregated to determine the total fixation duration in any particular response AOI.Since each participant was exposed to different interactions based on their response choices, few participants saw the same sequence or set of interactions.

Additionally, each interaction had varying numbers of response choices meaning that the total fixation times for any particular interaction was expected to vary significantly since participants would presumably require more time to consider more responses. In order to account for the variability between subjects, *percentage of fixation duration* was calculated for each interaction (Tsai et al. 2012).

The primary methods used to test each of the hypotheses were paired t-test and multiple linear regression.

6 RESULTS

This section organizes the results by each hypothesis.

6.1 H_{1a}: Participants will Spend Significantly More Time Fixating on Selected Answer Choices than Non-Selected Choices

To examine this hypothesis, four Bonferroni-adjusted post-hoc paired t-tests were conducted on the percentage of fixation duration. The first paired t-test conducted compared the percentage of fixation duration on the selected options and the combined fixation duration percentage of non chosen options. The second paired t-test conducted compared the percentage of fixation duration on the selected option and the average percentage of non chosen options. This set of paired t-tests was individually conducted on the data of each scene.

The paired t-test for the Library scene comparing the percentage of fixation duration on selected options(M=0.2246, SD=0.06645) and the combined duration percentage(M=0.3117, SD=0.1055) was significantly different(t=-4.2552, df=18, p=0.00048). The paired t-test for the Libary scene comparing the percentage of fixation duration on selected options(M=0.2426, SD=0.06645) and the average duration percentage(M=0.1232, SD=0.084) was significantly different(t=7.4842, df=18, p=6.247e-07).

The paired t-test for the Office Hours scene comparing the percentage of fixation duration on selected options(M=0.1985, SD=0.054) and the combined duration percentage(M=0.4224, SD=0.108) was significantly different(t=-7.613, df=18, p=4.929e-07). The paired t-test for the Office Hours scene comparing the percentage of fixation duration on selected options(M=0.1985, SD=0.054) and the average duration percentage(M=0.12315, SD=0.0298) was significantly different(t=4.9635, df=18, p=0.0001005).

These results indicate that participants generally spent more time fixating on selected options than any other option.

6.2 H_{1b}: Higher scoring participants will spend significantly more time fixating on culturally appropriate responses

In general, participants spent more time fixating on the correct options($M_{Library}$=0.340, $M_{OfficeHours}$=0.346) rather than incorrect options($M_{Library}$=0.196, $M_{OfficeHours}$=0.275).

H_{1b} was analyzed through linear regression employing the average Total Fixation on high scores as the dependent variable and the game score as the independent variable. The average Total Fixation on high scores as defined previously was calculated by summing the fixations on high scores for a participant and

dividing by the number of interactions the participant was exposed to throughout the game play labelled in the model as "App Fix Lib" and "App Fix OH" in Table 1. This analysis was performed separately on the Library scene and the Office Hours scene.

The Library scene model did not show the Game Score as being a significant predictor for the average Total Fixation on high scores. The Office Hours scene did show the Game Score as being a significant predictor for the average Total Fixation on high scores. The positive coefficient in this model indicates that higher game scores in the Office Hours scene were correlated to higher fixation duration percentages on high score responses.

Table 1: Regression Analysis for H_{1b}.

	Dependent variable:	
	App_Fix_Lib	App_Fix_OH
Game_Score_Lib	0.006 (0.007)	
Game_Score_OH		0.008** (0.004)
Constant	0.202 (0.163)	0.166 (0.081)
Observations	19	19
Adjusted R^2	−0.015	0.186
Residual Std. Error (df = 17)	0.101	0.063
F Statistic (df = 1; 17)	0.738	5.118**
Note:		**p<0.05; ***p<0.01

6.3 H_{1c}: The Participants with Advanced Domain Knowledge or Metacognitive Ability will Spend Significantly More Time Fixating on Higher-Score Responses

For the purposes of this hypothesis, higher domain knowledge and metacognitive ability are characterized by a participant's CQ score. Specifically, domain knowledge is captured by the Cognitive factor of CQ and metacognitive ability is captured by the Metacognitive factor of CQ. Based on a regression analysis where the dependent variable was the total Average fixation duration percentage on appropriate answers, both CQ factors were not shown to be significant as shown in Table 2. Additionally, we analyzed the data to determine if the fixation duration percentage on inappropriate responses was correlated to the CQ factors, but this also did not produce statistically significant factors as shown in Table 2.

6.4 H_{2a}: The Participants with Advanced Domain Knowledge or Metacognitive Skill Will Score Significantly Higher on the ETMP than those Participants with Lower Domain Knowledge.

Again, domain knowledge and metacognitive skill were characterized by the specific CQ components described in H_{1c}. Not surprisingly, the Metacognitive and Cognitive component of CQ were not significant factors in both the Library and the Office Hours scene regression analysis as shown in Table 3. When an Aikeke Information Criterion (AIC) based Stepwise regression was conducted, the CQ factors were eliminated (Akaike 1974).

6.5 H_{2b}: The participants who scored higher in the traditional game scoring will also score significantly higher on the ETMP.

As hypothesized, participants' game scores in traditional scoring as described as "Game Score Lib"(t=2.884, df=12, p=0.01374) and "Game Score OH"(t=8.832, df=12, p=1.35e-06) in Table 3 were positively correlated

Table 2: Regression Results for H_{1c} for Appropriate and Inappropriate Fixation Duration Percentage.

	Dependent variable:			
	App_Fix_Lib	App_Fix_OH	Inapp_Fix_Lib	Inapp_Fix_OH
CQ_Metacognitive	0.006 (0.035)	−0.010 (0.024)	−0.014 (0.028)	−0.022 (0.023)
CQ_Cognitive	−0.018 (0.032)	0.026 (0.022)	0.002 (0.026)	−0.001 (0.021)
Constant	0.400* (0.201)	0.266* (0.136)	0.262 (0.160)	0.405*** (0.134)
Observations	19	19	19	19
R^2	0.018	0.082	0.016	0.068
Adjusted R^2	−0.104	−0.032	−0.106	−0.048
Residual Std. Error (df = 16)	0.106	0.071	0.084	0.070
F Statistic (df = 2; 16)	0.149	0.718	0.134	0.585
Note:				*p<0.1; **p<0.05; ***p<0.01

with the ETMP of both the Library and Office Hours scenes. The AIC-based Stepwise Regression also preserved these factors.

6.6 H_{2c}: The Participants Who Exhibited a Gaze Bias on their Selected Answers Will Score Significantly Higher on the ETMP.

Participants who exhibited higher fixation duration percentages on their selected answers scored higher on the ETMP. The factor "Avg answer fixation Lib"(t=2.884, df=12, p=0.01374) as shown in Table 3 represented the fixation duration percentage on selected answers for the Library scene. The factor "Avg answer fixation OH"(t=3.048, df=12, p=0.01034) as shown in Table 3 represented the fixation duration percentage on selected answers for the Office Hours scene.

Additionally, the fixation duration percentage on the non-selected answers was also analyzed in the same model as described above. As seen in Table 3 the variables "Avg nonanswer fixation Lib"(t=3.247, df=12, p=0.00699) and the "Avg nonanswer fixation OH"(t=4.291, df=12, p=0.00105) were also found to be significant.

6.7 H_{2d}: The Participants Who Exhibited a Gaze Bias on Appropriate Answers will Score Significantly Higher on the ETMP.

Participants who exhibited high fixation duration percentages on appropriate answers as defined in the results of H_{1b} also resulted in higher ETMP scores as characterized by "Appropriate Fix Lib" (t=1.846, df=12, p=0.08967) and "Appropriate Fix OH"(t=2.447, df=12, p=0.03076) in Table 3. Of these findings, "Appropriate Fix Lib" was found to be significant at the $p<0.1$ level which indicates a weak correlation and should be further investigated with a larger sample size experiment.

7 DISCUSSION

The findings from this study are organized by the research questions previously described.

The first research question investigates general visual behaviors with respect to social interaction serious games like the CCDTS. We found that participants did, in fact, fixate more on their selections rather than the non-selected choices. This is consistent with various other studies that found longer fixation durations on selected options in multiple-choice based evaluation systems (Tsai et al. 2012; Hegarty et al. 1992). It is important to highlight, however, the large standard deviations of these results which may be an artifact

of the varying numbers of responses for each set of dialogue options. This warrants future experiments with consistent quantities of options per interaction to investigate this hypothesis.

Table 3: Regression Analysis for Library and Office Hours ETMP.

	Dependent variable:			
	ETMP Library		ETMP Office Hours	
	Original	Stepwise-AIC	Original	Stepwise-AIC
CQ_Metacognitive	−0.217 (0.345)		−0.192 (0.178)	
CQ_Cognitive	0.005 (0.327)		0.179 (0.167)	
Avg_answer_fix_Lib	10.043* (5.525)	10.406* (5.062)		
Avg_nonanswer_fix_Lib	14.869*** (4.579)	15.202*** (4.273)		
App_Fix_Lib	9.612* (5.207)	9.208* (4.868)		
Game_Score_Lib	0.252** (0.088)	0.248*** (0.082)		
Avg_answer_fix_OH			12.597** (4.133)	11.712** (4.014)
Avg_nonanswer_fix_OH			8.743*** (2.037)	8.699*** (1.937)
App_Fix_OH			9.816** (4.011)	10.817** (3.831)
Game_Score_OH			0.316*** (0.036)	0.315*** (0.034)
Constant	−4.158 (2.630)	−5.283*** (1.674)	−5.280*** (1.252)	−5.547*** (0.823)
Observations	19	19	19	19
Adjusted R^2	0.904	0.914	0.962	0.963
Residual Std. Error	1.024 (df = 12)	0.966 (df = 14)	0.507 (df = 12)	0.502 (df = 14)
F Statistic	29.234***	49.098***	77.420***	118.384***
Degrees of Freedom	(df = 6; 12)	(df = 4; 14)	(df = 6; 12)	(df = 4; 14)

Note: *p<0.1; **p<0.05; ***p<0.01

Additionally, we found a positive correlation between higher scoring participants and fixation duration on the more appropriate options in the Office Hours scene but not the Library scene. An explanation of this may be that the the Library scene was played prior to the Office Hours scene and, therefore, the user was visually calibrating his behavior resulting in more varied visual patterns. Assuming the Eye-Mind Theory, we interpret this result as being that higher scoring participants spent more time considering more appropriate answer options (Just and Carpenter 1980) and less time considering less appropriate responses. That being the case, we postulate that concentrated visual and mental focus on appropriate responses rather than less appropriate responses more concretely indicates higher cultural competency. This then sets the foundation for the validation of the ETMP.

Our analysis did not show significant correlation between our measures of domain knowledge or metacognition represented by Cultural Intelligence and fixation on appropriate answers. Though this measure has shown promise in various independent studies, the validation of this measure continues to be an active area of research and this finding motivates future experiments with other 3C constructs that have also shown potential(Chiu et al. 2013; Matsumoto and Hwang 2013).

With respect to the second research question exploring the validity of the ETMP as a measure of game performance, we found significant correlation between the ETMP and the other performance-related factors considered. Specifically, the linear regression models in Table 3 revealed significant correlations to how much of their relative attention was focused on the selected and non-selected options listed as "Avg answer fixation" and "Avg nonanswer fixation", respectively. This specific finding possibly indicates that the ETMP takes into account the validity of the Eye-Mind Theory (Just and Carpenter 1980). Additionally, we also found a correlation between the ETMP and fixation on the appropriate answers in the game as indicated by the "Appropriate Fix" factor. This can be interpreted to mean that the ETMP also considers the

accuracy of the mental schema that a player uses to determine his or her response though this interpretation is not entirely definitive. We also considered the possibility that rather this correlation may have occurred due to the low-score options being obviously socially unacceptable and easy to discard from consideration. This would then mean that the correct fixation was due in part to an understanding of general social norms rather than cultural competency. A targeted experiment with dialogue option sets that were more closely clustered in terms of appropriateness could resolve this interpretation. Finally, the ETMP was found to be strongly correlated to the player's explicit game score which we interpret as being that the players' explicit performances were likely attributable to cultural proficiency.

8 LIMITATIONS

The CCDTS dialogue structures were meant to facilitate a multitude of conversation tracks depending on the responses selected by the participants. Though this feature is ideal when considering realistic gameplay, it added complexity to the data structures and subsequently limited the sequential analysis that could be conducted since each player experienced different branches of the dialogue trees. For this reason, we were only able to effectively analyze the aggregated visual behavior metrics of each participant though we believe that a sequential analysis would provide insight into the specific metacognitive strategies employed by participants as well as performance improvement rates.

Also, despite instructions to maintain visual focus in the viewing area of the computer screen, several participants were observed consistently looking outside the limitations of the eye-tracker resulting in larger than expected track losses. Future experiments may warrant head-mounted devices in order to maintain tracking throughout the experiment.

Additionally, given the construct of a linear storyline of the game it was impossible to introduce the two scenes in varying orders. As such it is plausible that there may exist some findings that are dependent on the scene such as the correlation to the appropriate responses as observed in Hypothesis H_{1b}. Future experiments could involve the development of games that were not dependent on this sequencing in order to isolate the results from scene order.

Finally, the stimuli used in this game and the experimental design were dependent upon the game construct. Specifically, the CCDTS was designed as a DGBL to teach and assess cross cultural competence to individuals with a baseline knowledge of the Mandarin language and Chinese culture and, as such, limited the available participant pool. Future experiments should target DGBL systems that allow for a broader range of participants to be able to generalize the findings beyond the specific application presented in this study.

9 CONCLUSION

In summary, this study introduces a novel eye-tracking measure of performance in social interaction serious games. We present a unique methodology by which players of DGBL systems can be assessed for not only their explicit in-game responses, but also consider the mental schema used to determine the selected response. This study also presents the results of an experiment whereby the ETMP was found to be correlated to gaze-behaviors indicative of specific cognitive processes and traditional measures of in-game performance. These findings contribute to the initial validation of the ETMP and motivate further research of the ETMP in other DGBL domains.

REFERENCES

Adkins, S. S. 2017. "The 2017-2022 Global Game-based Learning Market". Technical report, Serious Play Conference 2017, Manassas, VA.

Akaike, H. 1974. "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control* 19(6):716–723.

Bayer, C., J. Jellá, A. Kapoor, R. Matts, H. Murdoch, H. Woodruff, B. An, S. Guerlain, and D. Brown. 2017. "The SIMPLE Methodology for Developing and Evaluating Cross-Cultural Virtual Training Systems SIEDS 2017". In *Systems and Information Engineering Design Symposium (SIEDS), 2017*, 289–293. IEEE.

Bellotti, F., B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta. 2013. "Assessment In and Of Serious Games: An Overview". *Advances in Human-Computer Interaction* 2013:1.

Bulling, A., J. A. Ward, H. Gellersen, and G. Troster. 2011. "Eye Movement Analysis for Activity Recognition Using Electrooculography". *IEEE transactions on pattern analysis and machine intelligence* 33(4):741–753.

Chiu, C.-Y., W. J. Lonner, D. Matsumoto, and C. Ward. 2013. "Cross-Cultural Competence: Theory, Research, and Application".

Coffey, A. J., R. Kamhawi, P. Fishwick, and J. Henderson. 2017. "The efficacy of an immersive 3D virtual versus 2D web environment in intercultural sensitivity acquisition". *Educational Technology Research and Development* 65(2):455–479.

Connolly, T. M., E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle. 2012. "A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games". *Computers & Education* 59(2):661–686.

Cook, T. D., D. T. Campbell, and W. Shadish. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd ed. Houghton Mifflin Boston.

Courtemanche, F., E. Aïmeur, A. Dufresne, M. Najjar, and F. Mpondo. 2011. "Activity Recognition Using Eye-Gaze Movements and Traditional Interactions". *Interacting with Computers* 23(3):202–213.

D'Mello, S., A. Olney, C. Williams, and P. Hays. 2012. "Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System". *International Journal of human-computer studies* 70(5):377–398.

Duchowski, A. T. 2002. "A Breadth-First Survey of Eye-Tracking Applications". *Behavior Research Methods, Instruments, & Computers* 34(4):455–470.

Fishwick, P. A., A. J. Coffey, R. Kamhawi, and J. Henderson. 2010. "An Experimental Design and Preliminary Results for a Cultural Training System Simulation". In *In the Proceedings of the 2010 Winter Simulation Conference (WSC)*, edited by B.Johannson et al., 799–810, Piscataway, New Jersey:IEEE.

Frutos-Pascual, M., and B. Garcia-Zapirain. 2015. "Assessing Visual Attention using Eye Tracking Sensors in Intelligent Cognitive Therapies based on Serious Games". *Sensors* 15(5):11092–11117.

Glaholt, M. G., and E. M. Reingold. 2011. "Eye Movement Monitoring as a Process Tracing Methodology in Decision Making Research.". *Journal of Neuroscience, Psychology, and Economics* 4(2):125.

Hegarty, M., R. E. Mayer, and C. E. Green. 1992. "Comprehension of Arithmetic Word Problems: Evidence from Students' Eye Fixations.". *Journal of Educational Psychology* 84(1):76.

Jaques, N., C. Conati, J. M. Harley, and R. Azevedo. 2014. "Predicting affect from gaze data during interaction with an intelligent tutoring system". In *International Conference on Intelligent Tutoring Systems*, 29–38. Springer.

Józsa, E., and B. Hámornik. 2011. "Find the Difference!: Eye Tracking Study on Information Seeking Behavior Using an Online Game".

Just, M. A., and P. A. Carpenter. 1980. "A Theory of Reading: From Eye Fixations to Comprehension.". *Psychological review* 87(4):329.

Kickmeier-Rust, M. D., E. Hillemann, and D. Albert. 2011. "Tracking the UFOs paths: Using Eye-Tracking for the Evaluation of Serious Games". In *International Conference on Virtual and Mixed Reality*, 315–324. Springer.

Kiili, K., H. Ketamo, and M. D. Kickmeier-Rust. 2014. "Evaluating the Usefulness of Eye Tracking in Game-based Learning". *International Journal of Serious Games* 1(2).

Lane, H. C., M. J. Hays, M. G. Core, and D. Auerbach. 2013. "Learning Intercultural Communication Skills with Virtual Humans: Feedback and Fidelity.". *Journal of Educational Psychology* 105(4):1026.

Lohse, G. L., and E. J. Johnson. 1996. "A Comparison of Two Process Tracing Methods for Choice Tasks". *Organizational Behavior and Human Decision Processes* 68(1):28–43.

Matsumoto, D., and H. C. Hwang. 2013. "Assessing Cross-Cultural Competence: A Review of Available Tests". *Journal of cross-cultural psychology* 44(6):849–873.

Renshaw, T., R. Stevens, and P. D. Denton. 2009. "Towards Understanding Engagement in Games: An Eye-Tracking Study". *On the Horizon* 17(4):408–420.

Shimojo, S., C. Simion, E. Shimojo, and C. Scheier. 2003. "Gaze Bias both Reflects and Influences Preference". *Nature neuroscience* 6(12):1317–1322.

Subrahmanyam, K., and P. M. Greenfield. 1994. "Effect of Video Game Practice on Spatial Skills in Girls and Boys". *Journal of applied developmental psychology* 15(1):13–32.

Tsai, M.-J., H.-T. Hou, M.-L. Lai, W.-Y. Liu, and F.-Y. Yang. 2012. "Visual Attention for Solving Multiple-Choice Science Problem: An Eye-Tracking Analysis". *Computers & Education* 58(1):375–385.

Van Dyne, L., S. Ang, and C. Koh. 2008. "Development and Validation of the CQS". *Handbook of cultural intelligence*:16–40.

Wang, Q., S. Yang, M. Liu, Z. Cao, and Q. Ma. 2014. "An Eye-Tracking Study of Website Complexity from Cognitive Load Perspective". *Decision support systems* 62:1–10.

Wierda, S. M., H. van Rijn, N. A. Taatgen, and S. Martens. 2012. "Pupil Dilation Deconvolution Reveals the Dynamics of Attention at High Temporal Resolution". *Proceedings of the National Academy of Sciences* 109(22):8456–8460.

Wu, C.-H., Y.-L. Tzeng, and R. Y. M. Huang. 2014. "A Conceptual Framework for Using the Affective Computing Techniques to Evaluate the Outcome of Digital Game-Based Learning". In *Advanced Technologies, Embedded and Multimedia for Human-centric Computing*, 189–196. Springer.

AUTHOR BIOGRAPHIES

BRIAN AN is currently a Ph.D. candidate in the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA. His e-mail address is baa4cb@virginia.edu.

INKI KIM is an Assistant Professor of Systems and Information Engineering at the University of Virginia. He holds a PhD in Industrial Engineering from the Pennsylvania State University. His research interests are in human behavior & performance modeling and simulation-based training. In particular, design, analysis, and improvement of "human(s)-in-the-loop" systems that rely on streamlined human interaction with highly-automated machines are common theme of his research. His e-mail address is ik3r@virginia.edu.

ERFAN PAKDAMANIAN is a Ph.D. student in the Department of System and Information Engineering at the University of Virginia. He received his M.S. in Industrial Engineering from Montana State University. His email address is ep2ca@virginia.edu.

DONALD E. BROWN is Director of the Data Science Institute at the University of Virginia and William Stansfield Calcott Professor of Engineering and Applied Science. His research focuses on data fusion, predictive models, and simulation optimization. His email address is deb@virginia.edu.