Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data

Article *in* International Journal of Human-Computer Interaction · April 2018

DOI: 10.1080/10447318.2018.1464776

6 authors, including:

Jiahui Wang
Kent State University
12 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Pavlo "Pasha" Antonenko
University of Florida
45 PUBLICATIONS **486** CITATIONS
SEE PROFILE

Yerika Jimenez
University of Florida
12 PUBLICATIONS **20** CITATIONS
SEE PROFILE

International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: http://www.tandfonline.com/loi/hihc20

Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data

Jiahui Wang, Pavlo Antonenko, Mehmet Celepkolu, Yerika Jimenez, Ethan Fieldman & Ashley Fieldman

To cite this article: Jiahui Wang, Pavlo Antonenko, Mehmet Celepkolu, Yerika Jimenez, Ethan Fieldman & Ashley Fieldman (2018): Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2018.1464776

To link to this article: https://doi.org/10.1080/10447318.2018.1464776

Published online: 30 Apr 2018.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2018.1464776

Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data

Jiahui Wang [a], Pavlo Antonenko[a], Mehmet Celepkolu[b], Yerika Jimenez[b], Ethan Fieldman[c], and Ashley Fieldman[c]

[a]School of Teaching and Learning, College of Education, University of Florida, Gainesville, FL, 32611, USA; [b]Department of Computer & Information Science & Engineering, College of Engineering, University of Florida, Gainesville, FL , 32611, USA; [c]Study Edge Corporation, Gainesville, FL, 32603, USA

ABSTRACT

This study explored the relationships between eye tracking and traditional usability testing data in the context of analyzing the usability of Algebra Nation™, an online system for learning mathematics used by hundreds of thousands of students. Thirty-five undergraduate students (20 females) completed seven usability tasks in the Algebra Nation™ online learning environment. The participants were asked to log in, select an instructor for the instructional video, post a question on the collaborative wall, search for an explanation of a mathematics concept on the wall, find information relating to Karma Points (an incentive for engagement and learning), and watch two instructional videos of varied content difficulty. Participants' eye movements (fixations and saccades) were simultaneously recorded by an eye tracker. Usability testing software was used to capture all participants' interactions with the system, task completion time, and task difficulty ratings. Upon finishing the usability tasks, participants completed the System Usability Scale. Important relationships were identified between the eye movement metrics and traditional usability testing metrics such as task difficulty rating and completion time. Eye tracking data were investigated quantitatively using aggregated fixation maps, and qualitative examination was performed on video replay of participants' fixation behavior. Augmenting the traditional usability testing methods, eye movement analysis provided additional insights regarding revisions to the interface elements associated with these usability tasks.

1. Introduction

With the exponential growth of information and communication technologies and their widespread application to promote formal and informal learning, online learning platforms have gained wide acceptance among teachers and students. Thus, understanding the user experience in these massive online learning systems is becoming increasingly more important. Comprehensive usability studies are essential in informing design and refinement of online learning systems and interfaces to improve the user experience. So far, a limited number of usability studies have been carried out to examine these massive online learning systems (Hasan, 2014; Ssemugabi & De Villiers, 2007).

The current study adopted a number of measures to comprehensively evaluate the usability of the Algebra Nation™, a massive online community for learning mathematics that is used by over 250,000 middle and high school students in the United States. From the methodological perspective, we focused on exploring the relationships between data generated using traditional usability testing techniques such as task difficulty ratings and eye-movement analysis data. Algebra Nation™ was designed to help students advance Algebra knowledge and skills and improve performance on the final Algebra exam to get high school graduation credit. The system offers instructional materials and support for students in areas including pre-Algebra, Algebra, and Geometry. For example, within the Algebra domain, Algebra Nation™ has a content review session where Algebra lessons are divided into 11 sections, and each of the sections contains 8–12 videos. The videos are designed as pencasts, where instructors write out the solution to a problem while explaining each step (Herold, Stahovich, Lin, & Calfee, 2011). Additionally, each video offers a picture-in-picture view of the instructor selected by the student from four different study experts that represent different races and genders. To reinforce the skills discussed in the video tutorials and provide a platform for social learning and peer support, Algebra Nation™ also provides an interactive collaborative wall where students can post questions about the material and get answers from their peers and study experts, and search for an answer in the existing threads. To encourage students to contribute to the interactive collaborative wall, Algebra Nation™ employs a rewards system called "Karma Points."

Usability studies of online learning technologies are still not common and little is understood about the usability aspects of massive online learning systems like Algebra Nation™. From a methodological standpoint, it is not always clear to usability researchers and practitioners when and why traditional usability testing methods like time on task and task

CONTACT Jiahui Wang jwang01@ufl.edu University of Florida, Norman Hall, 1221 SW 5th Avenue, Gainesville, FL 32611, USA
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

difficulty rating should be augmented with psychophysiological measures like eye tracking. This article addresses both issues by exploring the usability of Algebra Nation™ with a sample of its target users and by converging rigorous eye-movement analysis techniques and traditional usability testing methods to understand the user experience within a large online learning system. Several usability evaluation methods were adopted, including task completion time, task difficulty rating, System Usability Scale, and eye movement analysis including both gaze fixations and saccades. This study examined the relationships between these metrics relative to understanding the quality of user experience in an online learning system, which contributes to our understanding of these measures and their applicability in various contexts.

2. Literature review

Multiple usability testing methods are employed by usability evaluators and researchers to gather information on the quality of user experience. Widely used usability testing methods include, for example, measures of effectiveness (e.g., task success), efficiency (e.g., time on task), and satisfaction (International Organization for Standardization, 1998). With the advancement of sensing technologies and their availability to researchers and practitioners, eye tracking has gained popularity among usability scholars and professionals (Nielsen & Pernice, 2010). However, each usability testing method has its advantages and disadvantages, and to test a system and identify its usability problems, it is critical to select the most appropriate usability evaluation methods by considering the nature of human-system interactions being examined, the complexity of the system and interface, the time and cost involved in the usability testing, as well as the expertise of the usability evaluators.

2.1. Traditional usability testing methods

Usability is defined as the degree to which a product can be used by intended users to achieve specified goals with effectiveness, efficiency, and satisfaction (International Organization for Standardization, 1998). Following this widely adopted definition, usability performance concepts such as satisfaction, efficiency, and effectiveness (SEE) have been employed to design measures that assess whether and how a system is easy to use. Effectiveness has been measured by task success (i.e., the user's ability to complete the usability task successfully). Efficiency is typically assessed by how much time it takes the user to complete a usability task or the number of errors the user makes while completing the task. As task completion time and success do not necessarily capture all the elements associated with effectiveness and efficiency, studies have also elicited task difficulty ratings from users to measure effectiveness and efficiency (Tullis & Albert, 2013).

Satisfaction reflects the user's attitude or perceptions about system functionality and aesthetics and it has been measured using self-reports. Usability evaluators have employed standardized surveys to examine users' satisfaction with an interface (e.g., Everett, Byrne, & Greene, 2006). A number of standardized and validated usability surveys are available to measure participants' satisfaction such as the Computer System Usability Scale (CSUQ; Lewis, 1995), System Usability Scale (SUS; Brooke, 1996), and Questionnaire for User Interface Satisfaction (QUIS; Chin, Diehl, & Norman, 1988). In a systematic comparison of CSUQ, SUS, QUIS measures, SUS provided the most reliable results at sample sizes ranging from 6 to 14 (Tullis & Stetson, 2004). SUS, a highly robust measurement tool for usability researchers consistently producing reliable results (Bangor, Kortum, & Miller, 2008), has been adopted in usability testing of various products, ranging from everyday products such as microwave ovens (Kortum & Bangor, 2013) to mobile apps such as Gmail™ (Kortum & Sorber, 2015).

The traditional SEE usability metrics have been used in many usability studies (e.g., Rashid, Soo, Sivaji, Naeni, & Bahri, 2013). The three core aspects of usability—effectiveness, efficiency, and satisfaction—are equally important in usability testing, and they have been found to be highly dependent. For example, Kortum and Peres (2014) identified a strong positive correlation between task success rates (i.e., a measure of effectiveness) and SUS score (i.e., a measure of satisfaction), for both the laboratory and field studies at individual and system levels.

2.2. Eye tracking

Besides the traditional usability testing methods that focus on satisfaction, effectiveness, and efficiency, usability evaluators have started to adopt psychophysiological techniques to discover more insights about the user's attentional and cognitive processes during usability testing. Eye tracking in particular is a psychophysiological method that has recently gained much popularity among usability professionals. The main assumption behind the use of eye tracking in human factors and usability research is the eye-mind hypothesis (Just & Carpenter, 1980), which suggests that visual attention is the proxy for mental attention and so visual attention patterns reflect cognitive strategies used by individuals. Eye tracking has been employed to study visual attention distribution in a wide variety of visual tasks, from visual search (Pomplun, Reingold, & Shen, 2001) to reading (Schneps et al., 2013), viewing advertisements (Maughan, Gutnikov, & Stevens, 2007), to watching online video (Author, 2017). Eye tracking has also been applied in multiple usability studies to provide insights regarding the design of websites, digital TV menus, and games (Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007; Russell, 2005; Wulff, 2007).

Eye tracking has been a useful technique in user research, particularly in situations that require evaluation of the user's attention distribution relative to various (often competing) interface elements. With the recent advancements of sensor technology, eye tracking has also become more affordable and less intrusive to use (Pernice & Nielsen, 2009). However, to benefit from the information provided by eye tracking, one must understand specific eye movement metrics and what they represent.

Most modern eye trackers can accurately record two types of eye movements: gaze fixations and saccades (Rayner, 1998). A gaze fixation occurs when the eye focuses

on a visual target for a short period of time (i.e., around 300 ms). A saccade is a rapid eye movement between two fixations and saccades range in amplitude from small movements to large ones. Usability evaluators have examined these types of eye movement phenomena as quantified indices, such as the duration of each fixation, number of fixations, and saccade amplitude. For example, Wu and colleagues (2016) found that eye movement data such as fixation duration combined with fixation point number were useful in revealing how users search for target information on a smart watch interface, thus providing important information about interface information structure and interface element representation meaning. Also, Çöltekin and colleagues (2009) used eye tracking in a usability study of two online map websites, and the eye movement data including fixation durations and number of fixations revealed usability issues of specific features in two differently designed online map interfaces.

Eye tracking data can be examined not only quantitatively, but also qualitatively. Video replay of visual scan paths and eye movements can provide important insights regarding the patterns of attending to the various features of the interface, thus allowing usability researchers to identify where people focus their attention and for how long (Bojko, 2006; Goldberg & Kotval, 1999). For example, Wu and colleagues (2016) examined eye movement videos to describe the sequence of visual attention targets and identified potential usability issues within a smart watch interface. However, despite the potential to reveal important usability issues about an interface, eye tracking requires significant expertise and can be time consuming and labor intensive compared to traditional usability testing methods such as SEE measures (Pernice & Nielsen, 2009).

As each usability testing method has its advantages and disadvantages, usability researchers often adopt a combination of different usability testing techniques that complement one another (Jaspers, 2009; Tullis & Albert, 2013). An important gap in knowledge, however, is that the relationships between these usability methods for specific purposes and within specific contexts are not well understood. For example, it is not clear to usability researchers if and what specific eye tracking methods can provide added value to traditional measures of usability (Pernice & Nielsen, 2009), given that this method requires significant time, effort, specialized equipment, and expertise.

The current study adopted a number of usability methods to comprehensively evaluate the usability of the Algebra Nation™ massive online learning environment. The purpose of this study was to contribute to usability professionals' understanding of the relationships between eye movement metrics and the more traditional and easy-to-administer usability testing methods such as task completion time, task difficulty rating, and System Usability Scale.

3. Method

3.1. Participants

Thirty-five undergraduate students (ages 18–21; 20 females) who had never used Algebra Nation™ were recruited for this study. Approximately 74% of the participants identified themselves as White, and 20% were Hispanic. Participants represented multiple majors including finance, health science, international studies, psychology, and others. Twelve participants wore glasses or contact lenses (Table 1). None of the participants were color-blind.

Table 1. Participant demographics.

Variables	Statistics
Gender	20 female, 15 male
Age	M = 19.60 (SD = 0.85)
Ethnicity	26 White, 7 Hispanic, 2 Asian-Pacific Islander
Undergraduate classification	4 freshmen, 20 sophomores, 10 juniors, 1 senior
Wear glasses	12 Yes, 23 No

3.2. Usability tasks

Seven tasks that users typically complete within Algebra Nation™ were selected (see Table 2). Each task was designed to utilize one main feature of the Algebra Nation™ learning environment. Participants were instructed to complete these tasks to help researchers identify potential usability issues of the system.

Figures 1 through 4 show the Algebra Nation™ interface features participants used to complete the tasks. The final task focused on watching an instructional video either on Similar Triangles (easy topic) or Trigonometry (difficult topic). The video included a picture-in-picture effect of an instructor in the bottom right corner. The instructor explained the learning content using the typical non-verbal communication cues such as eye contact, facial expressions, and gestures. Figure 5 shows the screenshots of the two instructional videos on Similar Triangles (easy topic) and Trigonometry (difficult topic).

3.3. Apparatus

The Algebra Nation™ website was displayed on an external 20-inch flat screen monitor viewed at a 55-cm distance, at a 1600 by 1200 screen resolution and a 60 Hz refresh rate. Participants sat in a chair, and their head was stabilized using a chinrest built into the desk mount (SR Research, Ontario, Canada). An Eyelink 1000 Plus system and its Screen Recorder software were used to simultaneously capture locus of participants' eye movements and all screen activities,

Table 2. Tasks used in this study.

Task	Description
1	Find a way to log in to Algebra Nation™ using the following credentials. School: XXXXXX Username: xxxxxx Password: xxxxxx
2	Open Algebra I course, find the section on Quadratics—Part 1, and select the instructor you want to work with.
3	You are trying to solve the following equation. Seek help from the Algebra Nation™ community.$x^2 + \frac{x}{4} = 17$
4	Find a post explaining parallel lines.
5	Locate information about what Karma Points is.
6	Watch a video on Similar Triangles (easy topic).
7	Watch a video on Trigonometry (difficult topic).

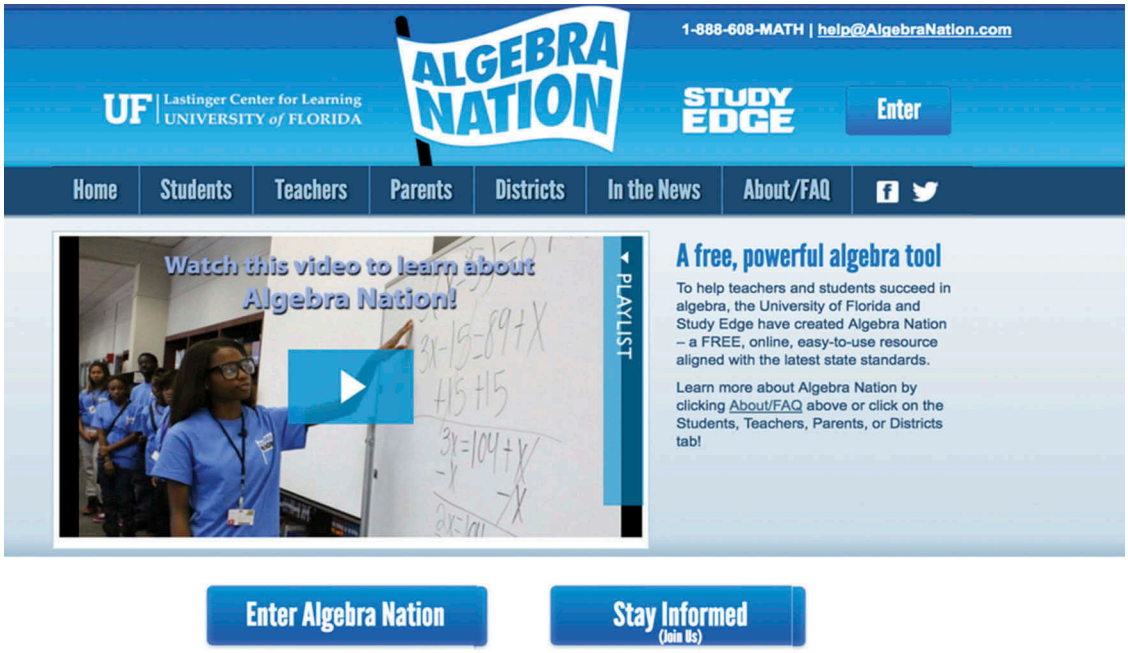

Figure 1. Task 1: Log in to the system.

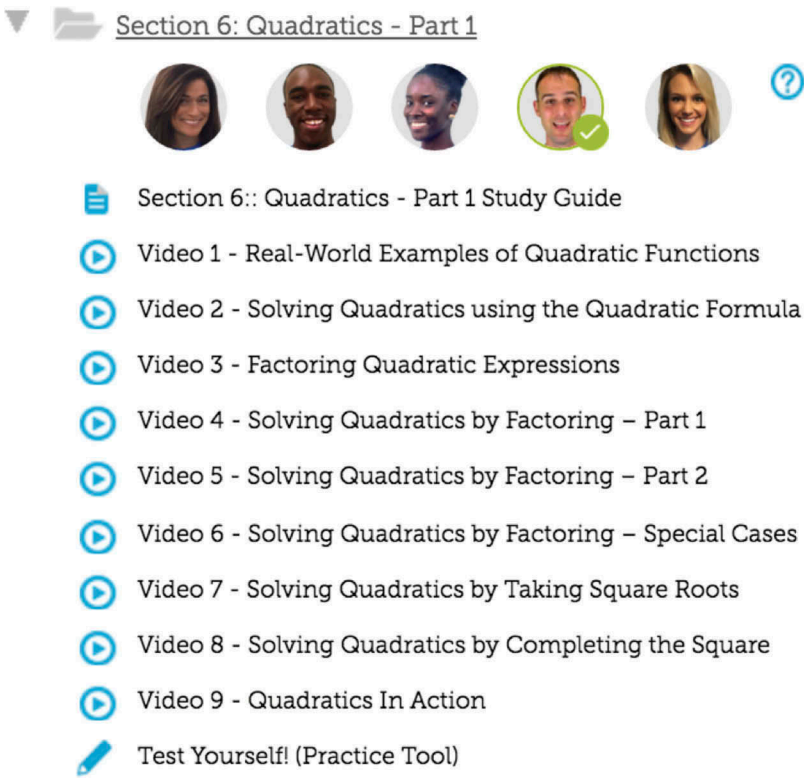

Figure 2. Task 2: Select an instructor.

as the participants performed the usability tasks (see Figure 6). Eye movements (i.e., fixations and saccades) were recorded at a sampling rate of 1000 Hz. Participants used a keyboard and a Bluetooth mouse as input devices.

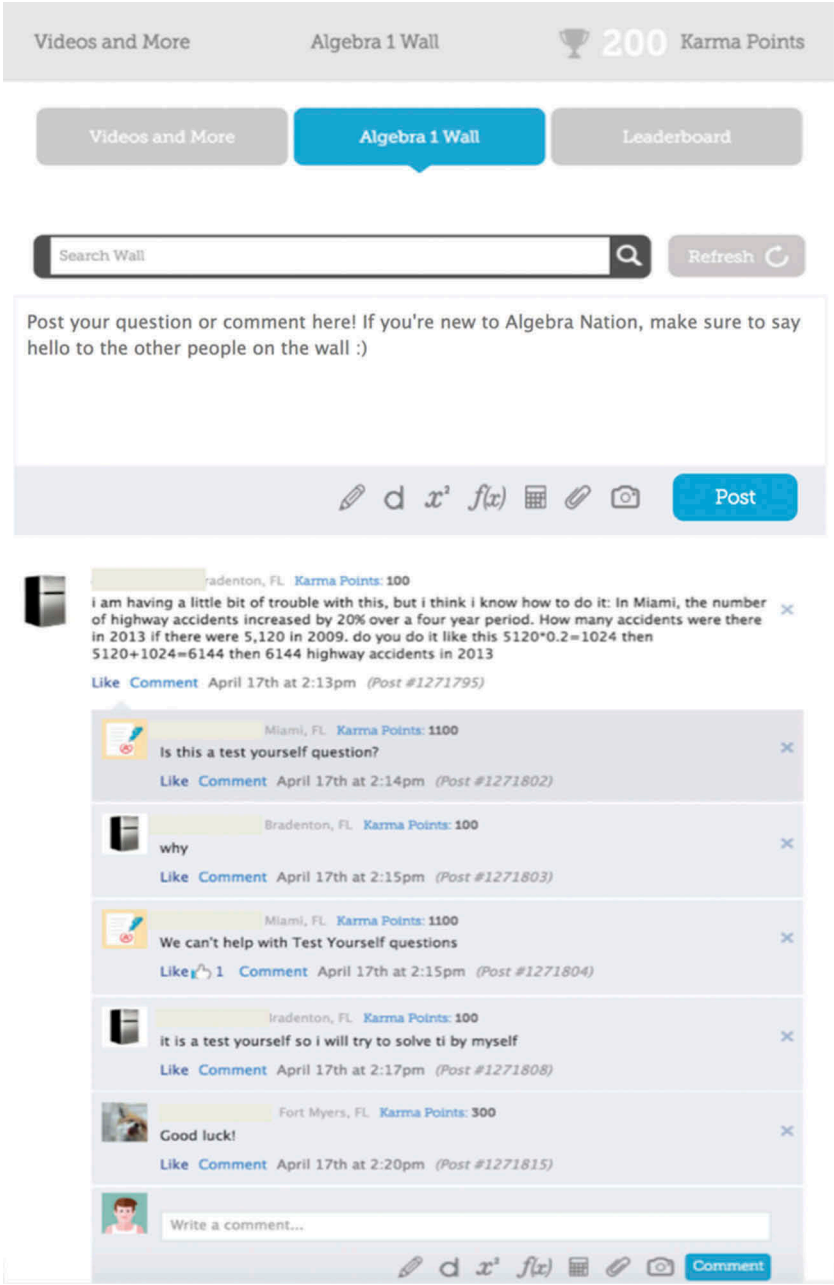

Figure 3. Task 3: Post an equation to the wall and task 4: find a post explaining parallel lines.

3.4. Procedure

After signing informed consent, participants completed a brief demographics survey. At the beginning of the experiment, the gaze of each participant was calibrated and validated with a 13-point calibration algorithm. Then, they were instructed to perform the usability tasks in Algebra Nation™. The instructions for each usability task were displayed in the middle top of the screen using TechSmith Morae™. Morae™ has been widely used for usability testing in various contexts (e.g., Çöltekin et al., 2009; Fagan, Mandernach, Nelson, Paulo, & Saunders, 2012). Morae™ recorded the completion time for each task and elicited post-task responses concerning task difficulty rating upon the completion of each task. Morae's™ task difficulty rating instrument is a five-point scale that ranges from very difficult (1) to very easy (5). While participants worked on each usability task and watched the instructional video, their eye movements (i.e., fixations and saccades) and on-screen activities were simultaneously recorded using Eyelink Screen Recorder. After completing the usability tasks, participants completed System Usability Scale (Brooke, 1996; see Table 3). Participants also reported their level of satisfaction with the video they watched on a nine-point scale that ranged from extremely dissatisfied (1) to extremely satisfied (9) (Author, 2017). The entire session lasted about 30 minutes for each participant.

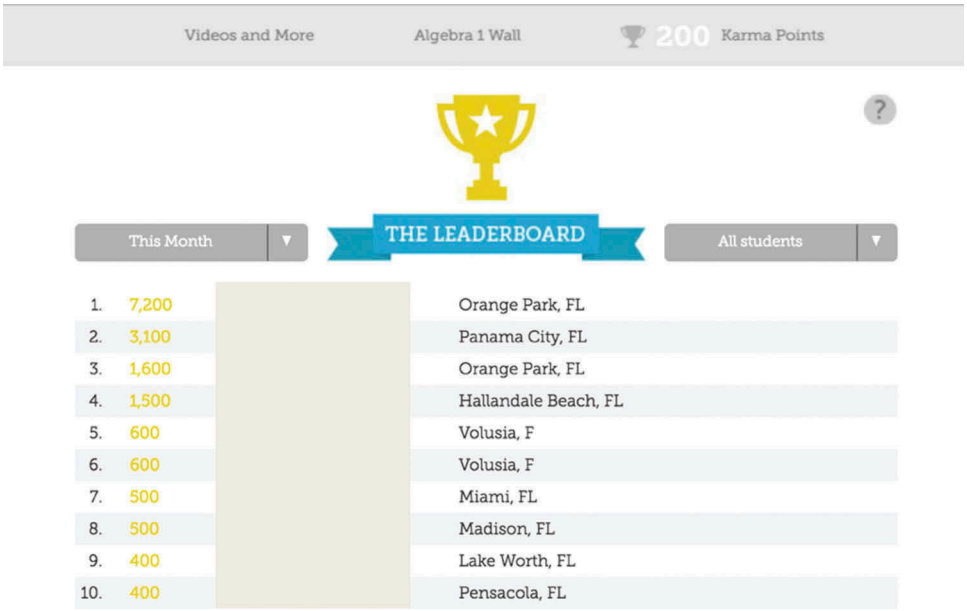

Figure 4. Task 5: Locate information about Karma Points.

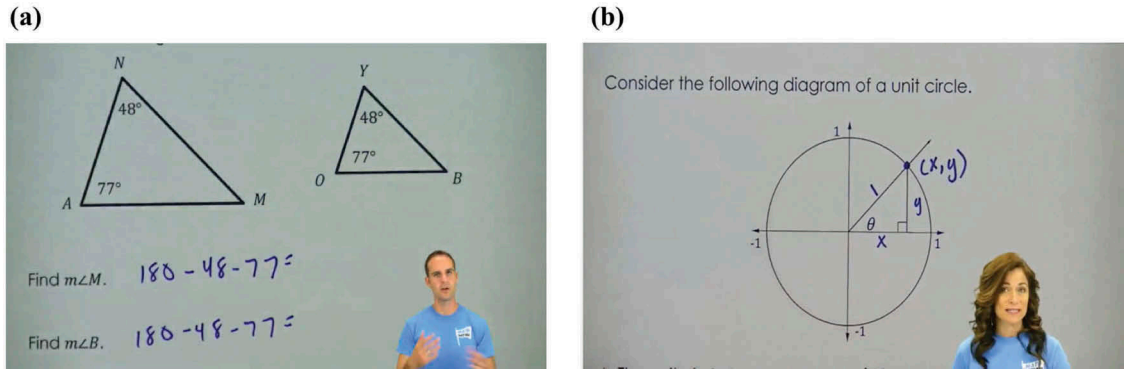

Figure 5. Tasks 6 and 7: Watch two instructional videos of varied content difficulty: (a) Similar triangles; (b) Trigonometry.

Figure 6. Experimental setup.

Table 3. System usability scale means for each item.

Statements	Mean
1: Strongly disagree to 5: Strongly agree	
Q1. I think that I would like to use this system frequently.	4.00
Q2. I found the system unnecessarily complex.	1.94
Q3. I thought the system was easy to use.	4.29
Q4. I think that I would need the support of a technical person to be able to use this system.	1.34
Q5. I found the various functions in this system were well integrated.	4.00
Q6. I thought there was too much inconsistency in this system.	1.37
Q7. I would imagine that most people would learn to use this system very quickly.	4.49
Q8. I found the system very cumbersome to use.	1.74
Q9. I felt very confident using the system.	4.11
Q10. I needed to learn a lot of things before I could get going with this system.	1.63

4. Results

The dependent variables measured by the traditional usability testing methods included task completion time, self-reported task difficulty rating, and System Usability Scale ratings. These data were complemented by eye movement data including the number of fixations, average fixation

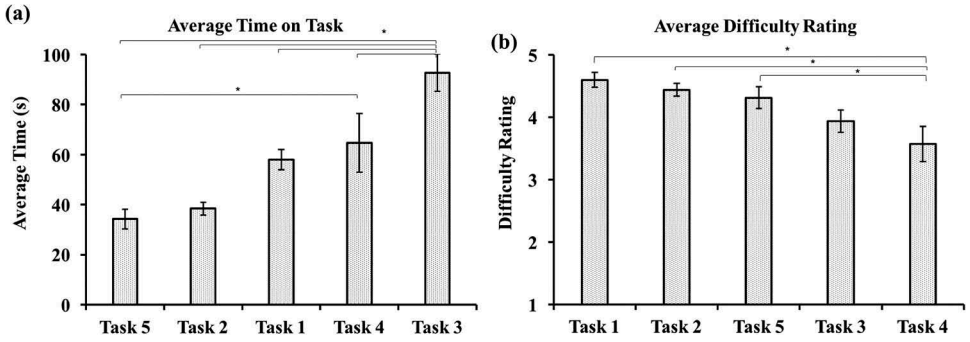

Figure 7. Average time (a) and task difficulty rating (b) for usability tasks. Rating: 1: very difficult, 5: very easy.
*means significant difference between two tasks. Error bars represent ± 1 SEM.

duration, and saccade amplitude for each usability task. Eyelink Data Viewer software (SR Research, Ontario, Canada) was used to divide the eye movement data into segments (one for each of the usability tasks) and extract the number of fixations, fixation duration, and saccade amplitude data for each task.

4.1. Task difficulty ratings and completion times

The average task difficulty rating for each task and average time spent on task are provided in Figure 7. ANOVA results indicated a significant difference in the task difficulty rating for the tasks (F (4, 168) = 5.17, $p < .05$, $\eta 2 = .11$). Bonferroni post-hoc analyses indicated participants rated task 4 ($\bar{X}$ = 3.57) as significantly more difficult compared to task 1 ($\bar{X}$ = 4.60), task 2 ($\bar{X}$ = 4.44), and task 5 ($\bar{X}$ = 4.31). Task 4 focused on finding a post explaining parallel lines. There was also a significant difference in task completion time (F (4, 168) = 11.951, $p < .05$, $\eta 2 = .222$). Participants took significantly longer to complete task 3 ($\bar{X}$ = 93s), as compared to task 1 ($\bar{X}$ = 58.2s), task 2 ($\bar{X}$ = 38.4s), task 4 ($\bar{X}$ = 64.8s), and task 5 ($\bar{X}$ = 34.2s). Task 3 focused on seeking help from the Algebra Nation™ community to help solve an equation. A significant difference was also found between the completion times for task 4 ($\bar{X}$ = 64.8s) and task 5 ($\bar{X}$ = 34.2s).

4.2. System usability scale (SUS) and satisfaction with the videos

Average overall System Usability Scale (SUS) score for all participants was 82, calculated following the method described in Brooke (1996), where the minimum is 0 and the maximum is 100. A higher score indicates a higher usability rating. A score of 82 represents "acceptable" usability (Bangor, Kortum, & Miller, 2009) and it is designated an A according to the Sauro-Lewis curved grading scale (CGS) for the SUS (Lewis & Sauro, 2017; Sauro & Lewis, 2016, p. 204). The mean score for each statement of SUS is reported in Table 3.

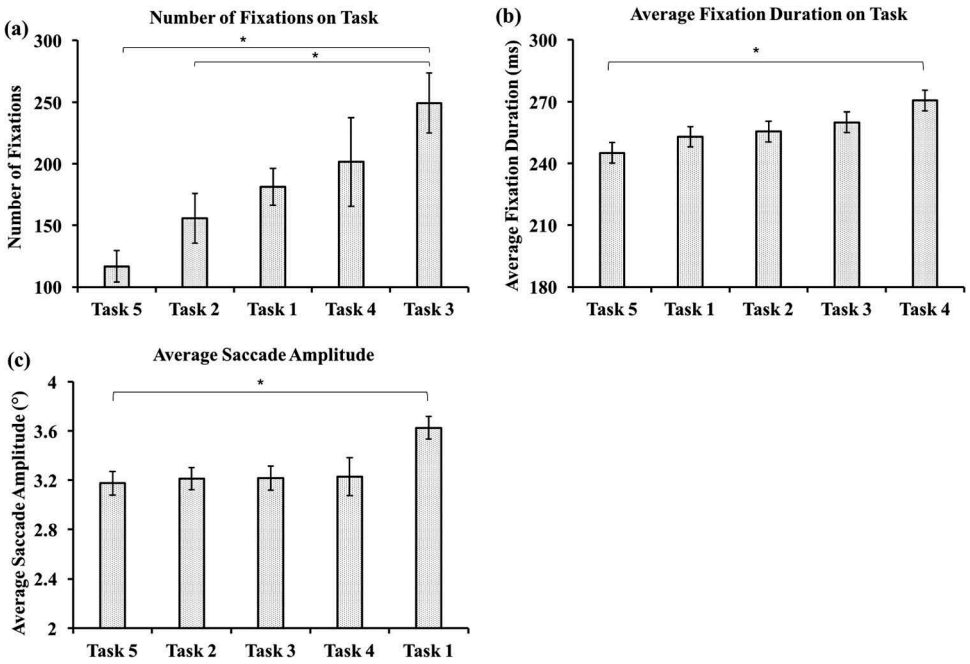

Figure 8. Number of fixations for each task (a), average fixation duration for each task (b), and average saccade amplitude for each task (c).
*means significant difference between two tasks. Error bars represent +/− 1 SEM.

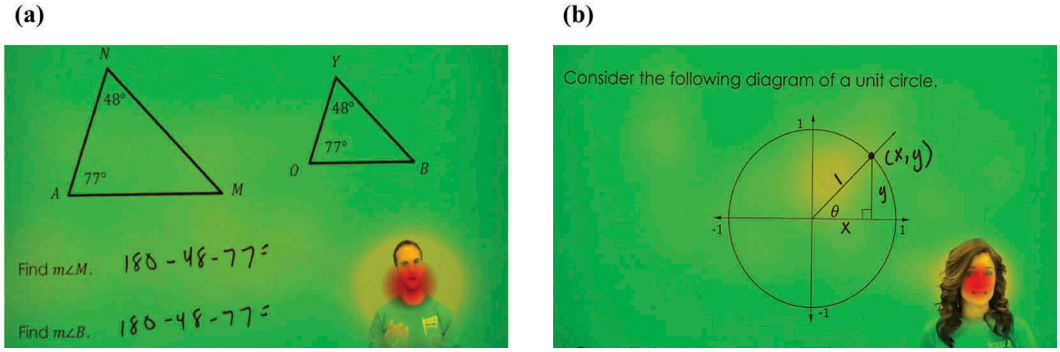

Figure 9. Aggregated heat maps of fixations on the two instructional videos of varied content difficulty: Similar triangles (a); Trigonometry (b).

For the videos participants watched, they reported a high level of satisfaction on a nine-point scale, 8.11 (SD = 0.68) for the easy topic on Similar Triangles and 7.61 (SD = 1.58) for the difficult video on Trigonometry.

4.3. Quantitative eye tracking data

We examined participants' number of fixations, average fixation duration, and average saccade amplitude for each usability task (see Figure 8). Saccade amplitude refers to the average size of saccades in degrees of visual angle. ANOVA results indicated a significant difference in the number of fixations (F (4, 168) = 4.690, p < .05, $\eta 2$ = .100), average fixation duration (F (4, 168) = 3.204, p < .05, $\eta 2$ = .071), and average saccade amplitude (F (4, 168) = 2.971, p < .05, $\eta 2$ = .066). Bonferroni post-hoc analyses indicated participants performed a significantly higher number of fixations during task 3 ($\bar{X}$= 249.23) compared to task 2 ($\bar{X}$ = 155.83) and task 5 ($\bar{X}$ = 116.71). Participants' average fixation duration was also longer when working on task 4 ($\bar{X}$ = 270.53 ms) compared to task 5 ($\bar{X}$ = 245.10 ms). Participants also produced significantly larger saccade amplitudes during Task 1 ($\bar{X}$ = 3.63°) compared to task 5 ($\bar{X}$ = 3.18°).

In addition to the five usability tasks, participants watched two instructional videos of varied content difficulty (i.e., Similar Triangles and Trigonometry). Figure 9 represents the aggregated fixation maps across the participants while they attended to the two videos. For the easy topic of similar triangles, the instructor attracted 26% of the total fixation time; whereas for the difficult topic of Trigonometry, the instructor attracted 22% of the total fixation time. Considering the instructor frame only constitutes 7% of the screen size, participants allocated a significant amount of visual attention to the instructor, especially the instructor's face. Participants generally expressed satisfaction with seeing the instructor on the screen and believed the instructor as helpful and engaging. These findings speak in favor of including the instructor in the Algebra Nation™ videos.

4.4. Qualitative eye tracking data

Besides analyzing the eye tracking data quantitatively, qualitative analysis was conducted on each participant's fixation behavior. The qualitative analysis provided information about where participants focused their attention and for how long, thus helping to identify potential usability problems within the system. Table 4 summarizes the usability problems associated with each of the usability tasks based on qualitative eye tracking data.

Table 4. Usability problems identified.

Usability Problem Description
Task 1: Log into the system
The "Enter" button on the top right corner of the main page was not immediately attended to by 6 participants.
Task 2: Select an instructor for the video
The question mark representing "about instructor" feature was not attended to or used by 34 participants.
Task 3: Post an equation to the wall
The equation editor sign$f(x)$ was not intuitive to the participant. Three participants confused it with the special character signx^2.
Task 4: Find a post explaining parallel line
The "refresh" button was very close to the "search" button on the search bar, and five participants clicked on the "refresh" button when they would like to search.
Two participants refreshed the page using the browser's refreshing functionality instead of using the "refresh" button on the search bar.
Search bar was not attended to by two participants, and they instead use control + F to search.
Task 5: Locate information about Karma Points
The question mark that led to information about Karma Points was not attended to by eight participants.

4.5. Relationships between usability metrics

Relationships between the usability metrics were examined using the Pearson product-moment correlation coefficient. We found a significant positive correlation between average task difficulty rating and the overall score for the System Usability Scale, r(33) = .768, p < .001. Higher SUS scores were associated with rating the tasks as easier to accomplish. Also, several significant positive and negative correlations were identified between task difficulty rating, task completion time, and eye movement metrics (i.e., number of fixations, average fixation duration, and average saccade amplitude). A summary of significant correlations is provided in Table 5.

We found that there was a moderate to strong negative correlation between the number of fixations and task difficulty rating for each task. Participants who exhibited more fixations during the task tended to rate the task as more difficult. We also identified a strong, positive correlation

Table 5. Significant correlations among usability metrics.

	Task 1	Task 2	Task 3	Task 4	Task 5
# of fixations and task difficulty rating	$r = -.334^{*}$	$r = -.484^{**}$	$r = -.432^{***}$	$r = -.622^{***}$	$r = -.650^{***}$
# of fixations and time on task	$r = .934^{***}$	$r = .646^{***}$	$r = .894^{***}$	$r = .972^{***}$	$r = .978^{***}$
Average saccade amplitude and time on task	/	/	$r = -.546^{***}$	/	$r = .415^{*}$
Average saccade amplitude and task difficulty rating	/	/	/	/	$r = -.424^{*}$
Time spent on task and task difficulty rating	/	$r = -.371^{*}$	$r = -.451^{**}$	$r = -.637^{***}$	$r = -.675^{***}$

Note: For task difficulty rating, 1: very difficult, 5: very easy.
$df = 33$.
$^{*}p < .05$.
$^{**}p < .01$.
$^{***}p < .001$.

between the number of fixations and time on task for each task. This means that participants who spent more time on a task also demonstrated more gaze fixations during that task. This finding was true for all tasks, although the strength of the relationship varied across the tasks.

Time on task was found to be negatively associated with task difficulty rating. Spending more time working on a task resulted in rating of the task as more difficult. This finding applied to all tasks except task 1 (log into the system) and for the other four tasks, the correlations ranged from strong (e.g., $r(33) = -.675$ for task 5, locate information on Karma Points) to moderate (e.g., $r(33) = -.371$ for task 2, select an instructor for the video). We have also identified a strong, negative correlation between average saccade amplitude and time on task for task 3 whereas a moderate, positive correlation between these two metrics for task 5. Moreover, average saccade amplitude and task difficulty rating were found to be negatively correlated for task 5. Task 5 required participants to locate information related to Karma Points and for this task, participants who demonstrated larger saccade amplitudes found the task more difficult.

Using average task difficulty rating as a predict variable, 59% of the variance in the SUS score is explained ($R^2 = .59$, $F(1, 33) = 47.398$, $p < .001$). Adding average fixation duration as a predictor variable in the model is associated with a statistically significant increase in R^2 ($\Delta R^2 = .048$, $F(1, 32) = 4.194$, $p < .05$). By using average fixation duration as a predictor, we can now predict 4.8% more variance in the SUS score than we could with a model that only contained average task difficulty rating.

5. Discussion

This study evaluated the usability of Algebra Nation™, a massive online learning environment that is used by hundreds of thousands of students, and investigated relationships between data collected using several usability evaluation methods. Traditional usability testing methods (i.e., standard metrics to gauge effectiveness and efficiency) revealed that the usability tasks resulted in variable task completion times and task difficulty ratings, which helped in identifying the aspects of the interface that need improvement. For example, participants rated task 4 (find a post explaining parallel line) as significantly more difficult compared to task 1 (log into the system), task 2 (select an instructor for the video) and task 5 (locate information on Karma Points). Results of the System Usability Scale, another traditional and widely used usability testing technique, suggested that Algebra Nation™ is user-friendly and easy to use. On average, the overall System Usability Scale score was 82, which is an acceptable SUS score for a system/interface evaluation. The levels of agreement with the SUS statements also corroborated this finding. Specifically, participants generally believed the system was easy to use and they were confident in using it.

In the current study, in addition to the traditional usability testing methods, eye movement metrics such as number of fixations, average fixation duration, and average saccade amplitude were examined. Eye tracking data results indicated that participants performed a significantly higher number of fixations during task 3 (seek help to solve an equation) compared to tasks 2 (select an instructor for the video) and 5 (locate information on Karma Points). Participants' average fixation duration was also longer when working on task 4 (find a post explaining parallel lines) compared to task 5 (locate information on Karma Points). Participants also produced significantly larger saccade amplitudes during task 1 compared to task 5. These results provide useful information about how different tasks induced different levels of visual attention from the participants and inform the aspects of the interface that can be improved.

Our study also examined the relationships between user's visual dynamics patterns collected using an eye tracker and these standard usability methods. These eye movement metrics reflected strong positive and negative correlations with task performance variables such as task completion time and task difficulty rating. First, in this study, negative correlations were identified between the number of fixations and self-reported task difficulty rating for each of the usability tasks. Specifically, higher number of fixations coincided with ratings representing higher levels of difficulty for each task. Importantly, this finding applied to all five usability tasks used in this study, ranging from a moderate correlation, $r(33) = -.334$ for task 1, log into the system) to a strong correlation, $r(33) = -.650$ for task 5, locate information on Karma Points). This finding confirms results of a study that used a visual search task in an experimental neuro-cognitive paradigm (Goldberg & Kotval, 1999). In that study, researchers evaluated several eye tracking measures that are relevant to the visual search task and suggested that when searching for a single target in a user interface, a larger number of fixations indicated that the user sampled many other objects prior to selecting the target. In other words, a larger number of fixations was associated with a less efficient visual search strategy due to less optimal interface layout. Based on these findings, it

is reasonable to conclude that in the current study, more fixations, possibly due to a suboptimal page layout, also resulted in the participants' rating those tasks as more difficult. Second, a series of positive correlations were identified between task difficulty rating and completion time. Rating a task as more difficult was positively associated with more time spent on the task. This association has been identified for all tasks except task 1 (log into the system), ranging from moderate correlation, $r(33) = .371$) to strong correlation, $r(33) = .675$). This finding is reasonable as users tend to spend more time figuring out how to complete a usability task when they perceive the task to be more difficult. Thus, it can be concluded that time on task can be used as a proxy for the difficulty in cognitive processing. This conclusion is consistent with the results reported in Cooke (2006), who found the easiest page resulted in the shortest task completion time.

In the current study, eye tracking metrics such as number of fixations and saccade amplitude provide convergent validity for the standard usability evaluation measures. Where eye tracking data provide a lot of added value is in discovering usability with individual interface elements that users attend to on the screen. Qualitative examination of eye tracking data can provide relatively unobtrusive measures of visual behavior that offer information about participants' attention and cognition, thus complementing the traditional usability evaluation methods in identifying usability issues at a deeper level. For example, by tracking eye movements, researchers were able to discover how long and how often a user looks at a certain area of interest in the interface and how frequently users switch from one visual component of the interface to others (Duchowski, 2007). In this study, the qualitative examination of eye tracking data provided us with detailed information regarding which features were used a lot or very little, thus leading to important insights on the solutions to the usability issues of the system (Table 6). The analysis especially suggested improving the design of the search bar in the Algebra Nation™ collaborative wall.

Before adopting eye tracking methods, usability researchers should consider the characteristics of the interface to be evaluated. Specifically, eye tracking could be useful in providing additional information about how users perceive different designs of an interface by examining the visual attention distribution over several areas of interest (AOIs). For example, in the current study, the aggregated fixation map is helpful in examining the users' visual attention distribution while they watched the two instructional videos which included the instructor on the screen. On the other hand, qualitative analysis of eye tracking data can provide valuable information on usability issues where users interact with interface that involves dynamically changing screens, for example, when the user is scrolling up and down a page to locate a piece of information.

Unlike most other research on massive online learning systems (Guo, 2013; Kiger, Herro, & Prunty, 2012), the current study focused on exploring the usability of the system, instead of simply examining the learning outcomes from the systems. Our study is one of the first few studies that used multiple evaluation methods to examine the usability of a massive online learning system. The eye tracking metrics such as number of fixations provide convergent validity for the standard usability evaluation measures. More importantly, the qualitative examination of eye movement data revealed several design flaws of the system and provided important suggestions on how to improve the interface design, which is otherwise impossible to acquire from using traditional usability testing methods.

Table 6. Solutions to usability problems.

Usability Problem Description	Solutions
Task 1: Log into the system	
The "Enter" button on the top right corner of the main page was not immediately attended to by 6 participants.	Use "Log in" instead of "Enter" on the main page. Use a contrasting color of blue for the "log in" button. Eliminate the "Enter" button in the center of the main page.
Task 2: Select an instructor for the video	Make the question mark for "about instructor" stand out more by using a different color and a more intuitive icon.
The question mark representing "about instructor" feature was not attended to or used by 34 participants.	Create a hover over feature with a short description about the instructors or create a separate page for instructors' information in the system.
Task 3: Post an equation to the wall	
The equation editor sign$f(x)$ was not intuitive to the participant. It confused three participants with the special characters signx^2.	Change the looks of special characters and equation editor signs and make them look more self-explanatory.
Task 4: Find a post explaining parallel line	
The "refresh" button was very close to the "search" button on the search bar, and five participants clicked on the "refresh" button when they would like to search.	Eliminate the "refresh" button.
Two participants refreshed the page using the browser's refreshing functionality instead of using the "refresh" button on the search bar.	Make the search stand out more by using a different color or assigning a bigger space.
Search bar was not attended to by two participants, and they instead use control + F to search.	
Task 5: Locate information about Karma Points	
The question mark that led to information about Karma Points was not attended to by eight participants.	Make the question mark stand out more by using a different color (e.g., blue). Create a hover over caption saying, "what is Karma points?" over the question mark.

6. Conclusion

his study explored the relationships between eye tracking data and standard usability testing data that focus on the effectiveness and efficiency of completing usability tasks. The context of the study was evaluating the usability and cognitive task requirements of Algebra Nation™, a massive online learning environment used by hundreds of thousands of students in the USA. The usability tasks resulted in variable levels of self-reported task difficulty rating and completion time, which helped identify the aspects of the interface that need improvement. Compared to traditional usability metrics that gather data based on participants'

overt behavior (e.g., time on task), qualitative eye movement analysis provided additional insights into how participants respond to visual elements while interacting with the interface. Eye movement analysis augmented the traditional usability testing methods, and provided important implications regarding revisions to the interface elements associated with those usability tasks.

Acknowledgment

We also would like to thank the Study Edge Corporation for their help with participant recruitment.

Funding

This article is based on work supported by the National Science Foundation under Grant No. 1540888.

ORCID

Jiahui Wang http://orcid.org/0000-0001-5681-5055

References

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4* (3), 114–123.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*(March), 574–594. doi:10.1080/10447310802205776

Bojko, A. (2006). Using eye tracking to compare web page designs: A case study. *Journal of Usability Studies*, *1*(3), 112–120.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of theSIGCHI conference on Human factors in computing systems* (pp. 213–218). ACM.

Çöltekin, A., Heil, B., Garlandini, S., & Fabrikant, S. I. (2009). Evaluating the effectiveness of interactive map interface designs: A case study integrating usability metrics with eye-movement analysis. *Cartography and Geographic Information Science*, *36*(1), 5–17. doi:10.1559/152304009787340197

Cooke, L. (2006). Is eye tracking the next step in usability testing?. *International professional communication conference, 2006 IEEE* (pp. 236–242). doi:10.1109/IPCC.2006.320355

Cowen, L., Ball, L. J., & Delin, J. (2002). An eye movement analysis of web page usability. In *People and computers XVI-memorable yet invisible* (pp. 317–335). London: Springer.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice*. London: Springer-Verlag.

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. *In 21st British CHI group annual conference on HCI 2007: People and Computers XXI, 1*, (p. 12). Swindon, UK: The British Computer Society. doi:10.1145/1531294.1531311

Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(24), 2547–2551. doi:10.1177/154193120605002407

Fagan, J. C., Mandernach, M. A., Nelson, C. S., Paulo, J. R., & Saunders, G. (2012). Usability test results for a discovery tool in an academic library. *Information Technology and Libraries*, *31*(1), 83

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.doi:10.1016/S0169-8141(98)00068-7

Guo, P. J. (2013). *Online python tutor: Embeddable web-based program visualization for cs education*. SIGCSE 2013 - Proceedings of the 44th ACM Technical Symposium on Computer Science Education. 579–584. doi:10.1145/2445196.2445368

Hasan, L. (2014). Evaluating the usability of educational websites based on students' preferences of design characteristics. *International Arab Journal of E-Technology*, *3*(3), 179–193.

Herold, J., Stahovich, T., Lin, H. L., & Calfee, R. C. (2011).The effectiveness of "pencasts" as an instructional medium. Proceedings of the American Society for Engineering Education 118th Annual Conference and Exposition.

Jaspers, M. W. M. (2009). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International Journal of Medical Informatics*, *78*(5), 340–353. doi:10.1016/j.ijmedinf.2008.10.002

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. doi:10.1037/0033-295X.87.4.329

Kiger, D., Herro, D., & Prunty, D. (2012). Examining the influence of a mobile learning intervention on third grade math achievement. *Journal of Research on Technology in Education*, *45*(1), 61–82. doi:10.1080/15391523.2012.10782597

Kortum, P., & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the system usability scale. *International Journal of Human–Computer Interaction*, *30*(7), 575–584. doi:10.1080/10447318.2014.904177

Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human-Computer Interaction*, *31*(8), 518–529. doi:10.1080/10447318.2015.1064658

Kortum, P. T., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. *International Journal of Human-Computer Interaction*, *29*(2), 67–76. doi:10.1080/10447318.2012.681221

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. doi:10.1080/10447319509526110

Lewis, J. R., & Sauro, J. (2017). Can I leave this one out? The effect of dropping an item from the SUS. *Journal of Usability Studies*, *13*, 1.

Maughan, L., Gutnikov, S., & Stevens, R. (2007). Like more, look more. Look more, like more: The evidence from eye-tracking. *Journal of Brand Management*, *14*(4), 335–342. doi:10.1057/palgrave.bm.2550074

Nelsen, B. T. (1998). *Ergonomic requirements for office work with visual display terminals part 11: Guidance on usability (ISO DIS 924–11)*. London: ISO.

Nielsen, J., & Pernice, K. (2010). *Eyetracking web usability*. Berkeley, CA: New Riders.

Pernice, K., & Nielsen, J. (2009). *How to conduct eyetracking studies*. Fremont, CA: Nielsen Norman Group.

Pomplun, M., Reingold, E. M., & Shen, J. (2001). Investigating the visual span in comparative search: The effects of task difficulty and divided attention. *Cognition*, *81*(2), 57–67. doi:10.1016/S0010-0277(01)00123-8

Rashid, S., Soo, S., Sivaji, A., Naeni, H. S., & Bahri, S. (2013). Preliminary usability testing with eye tracking and FCAT analysis on occupational safety and health websites. *Procedia - Social and Behavioral Sciences*, *97*, 737–744. doi:10.1016/j.sbspro.2013.10.295

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. doi:10.1037/0033-2909.124.3.372

Russell, M. (2005). Using eye-tracking data to understand first impressions of a website. *Usability News*, *7*(1), 1–14.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Cambridge, MA: Morgan Kaufmann.

Schneps, M. H., Thomson, J. M., Sonnert, G., Pomplun, M., Chen, C., & Heffner-Wong, A. (2013). Shorter lines facilitate reading in those who struggle. *PLoS ONE*, *8*(8). doi:10.1371/journal.pone.0071161

Ssemugabi, S., & De Villiers, R. (2007). *A comparative study of two usability evaluation methods using a web-based e-learning application*. Proceedings of the 2007 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on It Research in Developing Countries, (April 2016), 132–142. doi:10.1145/1292491.1292507

Tullis, T., & Albert, W. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed). Amsterdam: Morgan Kaufmann.

Tullis, T. S., & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. *Usability Professional Association Conference*, 1–12.

Wang, J., & Antonenko, P. (2017). Instructor presence in instructional video: Effects on visual attention, recall, and perceived learning. *Computers in Human Behavior*, *71*, 79–89.

Wu, Y., Cheng, J., & Kang, X. (2016). Study of smart watch interface usability evaluation based on eye-tracking. In *International conference of design, user experience, and usability* (pp. 98–109). Cham: Springer International Publishing. doi:10.1007/978-3-319-20886-2

Wulff, A. (2007). Eyes wide shut-or using eye tracking technique to test a website. *International Journal of Public Information Systems*, *3*(1), 1–12.

About the Authors

Jiahui Wang is a PhD candidate and research fellow in Educational Technology program at the University of Florida. Her research focuses on individual differences affect learning with technology and how technology-based learning environments can be designed to accommodate individual differences.

Pavlo "Pasha" Antonenko is an Associate Professor of Educational Technology and Director of NeurAL Lab at the University of Florida. His scholarship focuses on a) frameworks and technologies to encourage and scaffold learning and twenty-firstcentury skills and b) psychophysiological assessment of cognition to optimize design of technology-enhanced learning environments.

Mehmet Celepkolu is currently a PhD student and in the Computer Science program at the University of Florida. His research focuses on how computational models can reveal the hidden phenomena during dialogue and learning, which can create effective strategies for supporting human learning with intelligent systems.

Yerika Jimenez is currently a PhD student in Human-Centered Computing and an NSF graduate research fellow at the University of Florida. Her research focuses on computer science education. She is interested in understanding how much cognitive effort do students use to interact and learn computer science with block-based programming environments.

Ethan Fieldman is the President of Study Edge. Study Edge provides various services including education technology that uses social media, mobile devices, online communities, gamification, personalized learning, and some of the best, most energetic instructors in the world to help students from middle school through college.

Ashley Fieldman is the Vice President of Study Edge. Study Edge provides various services including education technology that uses social media, mobile devices, online communities, gamification, personalized learning, and some of the best, most energetic instructors in the world to help students from middle school through college.