Understanding Student Success in Chemistry using Gaze Tracking & Pupillometry

Joshua Peterson[1], Zachary Pardos[1], Martina Rau[2], Anna Swigart[1], Colin Gerber[1], and Jonathan McKinsey[1]

[1] University of California, Berkeley
{jpeterson,pardos}@berkeley.edu
[2] University of Wisconsin – Madison
marau@wisc.edu

Abstract. Eye tracking allows us to identify visual strategies through gaze behavior, which can help us understand how students process content. Furthermore, understanding which visual strategies are successful can help us improve educational materials that foster successful use of these visual strategies. Previous studies have demonstrated the predictive value of eye tracking for student performance. Chemistry is a highly visual domain, making it particularly appropriate to study visual strategies. Eye tracking also provides measures of pupil dilation that correlate with cognitive processes important to learning, but have not yet been assessed in any realistic learning environments. We examined the gaze behavior and pupil dilation of undergraduate students working with a specialized ITS for chemistry: Chem Tutor. Chem Tutor emphasizes visual learning by focusing specifically on graphical representations. We assessed the value of over 40 high-level gaze features along with measures of pupil diameter to predict student performance and learning gains across an entire chemistry problem set. We found that certain gaze features are strong predictors of performance, but less so of learning gains, while pupil diameter is marginally predictive of learning gains, but not performance. Further studies that assess pupil dilation with higher temporal precision will be necessary to draw conclusions about the limits of its predictive power.

Keywords: Eye tracking, intelligent tutoring systems, performance prediction, chem tutor

1 Introduction

Eye tracking provides behavioral and physiological metrics that researchers can use to study a number of psychological and physiological processes. In the context of education, these metrics can reveal visual strategies and provide clues as to how students process content. Armed with such knowledge, instructional designers can build better content and interfaces for Massive Open Online Courses (MOOCs), Intelligent Tutoring Systems (ITS), and with the advent of affordable and wireless head-mounted trackers, perhaps even traditional classrooms. While eye tracking research applied to education has already begun to yield insights into students'

behavior and internal states, we identify two important research questions that deserve considerable attention, namely, whether gaze behavior predicts performance and learning gains in a highly visual, STEM-related, domain-specialized ITS, and whether the recognized utility in measuring cognitive processes by tracking pupil diameter transfers to realistic learning contexts.

Current eye tracking technology provides information on blink rate, fixation, saccades and pupil diameter at high sampling rates [1]. Blink rate is a measure of how quickly eyelids are closed, then opened. Saccades are abrupt, rapid movements from one element to another. Fixation is the amount of time the eyes are focused on a given point on the screen, such as a letter within a word. Pupillometry is concerned with measurement of the pupillary diameter and its fluctuation over time in response to external stimuli or internal state changes.

Much of the current work in applying eye tracking to education has focused on what content students fixate on, for how long, and in what order or sequence. Some of this research has been aimed at distinguishing the gaze behavior of high-performing and low-performing students. For instance, when given a standardized multiple-choice science exam involving chemistry, biology and physics questions, participants who had more expertise in a specific subject area needed fewer eye fixations to process information in problem statement, image, and multiple choice zones and had fewer saccades between zones [2]. High-performing students also spend more time looking at relevant problem details and candidate solution choices than low-performing students [3]. A more recent study sought to understand the differences in eye tracking patterns between high and low performers in three engineering-related computer games that required spatial ability, problem-solving skills, and a capacity to interpret visual imagery [4]. Successful players showed shorter first fixations after a stimulus presentation, which has been correlated with high attentional readiness [1]. High performers also used fewer clicks, more unique fixation points, and a longer duration on average for each eye fixation, which was speculated to be associated with engagement and cognitive processing prior to taking action. In contrast, low performers were characterized by longer first fixations after stimulus presentation, more mouse clicks, and shorter durations for each fixation point, which might suggest a trial-and-error approach. This constitutes the first attempt to our knowledge to apply eye tracking to a highly visual problem-solving domain that focuses entirely on skills that are important to core subjects like chemistry [5]. One pathway to expanding such an endeavor will require a closer look at several aspects of gaze behavior in learning environments aimed specifically at more direct instruction of the target field of study.

The role of pupillometry in the science of learning and education is far less explored than gaze behavior. However, there is reason to believe it may provide equal insight to the field. In cognitive psychology, pupillometry has been shown to indicate a number of cognitive processes in highly-controlled cognitive task paradigms. Notably, pupil size correlates with the difficulty of a task across a number of domains [6]. For example, pupils dilate while doing difficult versus easy multiplication problems, recalling complicated sentences [7], and performing difficult analogy tasks [8]. Pupils also reliably dilate with the number of digits to be remembered, reflecting

short term memory load [6]. Lastly, pupil diameter has even been shown to fluctuate with attention and mental effort [9]. Pupillometry is a more accurate, less noisy measure of these processes than EEG and a cheaper and more practical alternative to imaging methods such as fMRI. Given the utility of pupillometry in indexing cognitive processes across a wide range of tasks, it is reasonable to assume that pupil dilation measurements may help predict student performance and learning gains. Further, our current understanding of cognitive load has led to improvements in instructional design and procedures that can improve learning [10]. Pupillometry may provide a new window into cognitive processes such as cognitive load during complex realistic learning scenarios, but only if the utility of which can be shown to be more externally valid. An intelligent tutoring system provides such a complex learning environment as compared to controlled cognitive tasks.

Here, we ask whether a large number of gaze features along with pupil diameter can predict student performance and learning gains in learning concepts from chemistry in an intelligent tutoring system.

2 Methodology

2.1 Experimental Design and Data Collection

We obtained gaze and pupil diameter for 95 undergraduate students using a SMI RED 250 eye-tracker as they worked with an ITS for chemistry: Chem Tutor [11]. This data set was drawn from an experiment that investigated the effects of different types of support students were given for making connections between graphical representations such as Lewis structures and ball-and-stick figures (please refer to [11] for information on these support types). Figure 1 shows a truncated example of the student-problem level data. Students had to complete all problems that were part of the intervention of the experiment. Each problem contained a series of steps with unlimited attempts. Students could request hints from Chem Tutor for steps they struggled with. In addition, each student took a pretest and posttest to assess reproduction and transfer of representational skills and chemistry knowledge.

student	**problem**	**errorRate_medianSplit**	**...**	**1stFixDur_GR**
1	U1_I1_1	0	...	347
1	U1_I2_1	0	...	203
1	U1_I3_1	1	...	115
1	U2_I1_1	1	...	320
...	...	...	...	...

Figure 1. Truncated example of data set

2.2 Feature Engineering

We constructed forty problem-level gaze features averaged over the time taken to complete each problem. From the raw eye tracking data, we created areas of interest (AOIs) from elements of the tutor that included graphic representations (GR), whitespace, titles, hints, the progress bar, and the interaction pane. We use "whitespace" here as a catch-all for any screen space that was not occupied by other AOIs. We computed frequency of switching between unique AOIs, a metric that has been associated with perceptual integration [12]. We also computed the frequency of switching between GRs as the number of times a fixation on one AOI was followed by one on another AOI. Next, we computed the duration of fixation after the first inspection of an AOI. A first inspection of an AOI is thought to indicate early processing of content [13]. The duration of fixations after the first inspection is thought to reflect intentional processing to integrate one source of information with another [13]. We then computed the duration of second-inspection fixations on each AOI as the sum of fixation durations that occurred after the first fixation on AOIs. In addition, we computed the sum of total fixation durations on each AOI. Beyond single-AOI fixation features, we also computed fixation sequence features. These features involve a specific ordering of fixation targets. For instance, a student may first focus on a graphic, then on the text of a hint, and finally back to the graphic again. The majority of them involve focus on GRs in reference to other information, computed as the counts of fixations on one AOI followed by fixation on either one other (2-point sequence), or two other AOIs (3-point sequence).

2.3 Analysis

To investigate which features were predictive of student performance and learning gains, we first identified a number of features that correlated with performance. Given that most of our features were not normally distributed and difficult to normalize through transformations, we calculated independent Pearson correlations between our outcome and predictors since they do not assume normality. Our performance outcome variable was a metric termed first-incorrect rate (FIR) that is defined as the number of times a student gave an incorrect answer on first attempt normalized by the total number of steps for that question, which is standard practice in ITS research [14]. We used the number of first-attempt incorrect instead of the total number of incorrect attempts in order to capture performance on new steps (students could re-attempt steps until they were correct). Our outcome variable representing a learning gains score (LGS) was defined as posttest minus pretest score for each student. To predict student performance and learning gains, we used logistic regression on a median split of the criterion variables. Due to the issues with non-normal features, our final model used Gaussian logistic regression on a median split instead of multiple regression on a continuous performance criterion. Finally, we evaluated each model using four-fold cross validation.

3 Results

Several fixation features significantly predicted FIR (see Table 1), and were highest when averaged over problems. Correlations between FIR and all other features including pupil diameter were not significant. The majority of these features involved fixation on titles, the progress indicator, and the interaction pane. All associations were positive (e.g. longer fixation on titles accompanied higher error rates) with the exception of those involving the progress indicator and screen/interface whitespace. Fixation on titles seemed to have the strongest negative effect on error rate. This may be because lower performers tended to focus longer on the titles when they were confused about the topic. Features involving fixation on the progress indicator were consistently associated with smaller error rates, perhaps because learners who kept track of their progress to budget time or used progress as motivation perform better as a result.

Table 1. Correlations between FIR (first-incorrect rate) and selected fixation features

Feature	***r***	**Feature Description**
fixDur_Titles	0.80***	sum of fixation durations on titles
fixDur_Progress	-0.61**	sum of fixation durations on progress indicator
fixDur_Interaction	0.48*	sum of fixation durations on interaction pane
1stFixDur_Titles	0.81***	duration of first fixation on titles
1stFixDur_Interaction	0.60**	duration of first fixation on interaction pane
1stFixDur_Progress	-0.43*	duration of first fixation on progress indicator
1stFixDur_WhiteSpace	-0.50*	duration of first fixation on whitespace
2ndFixDur_Titles	0.80***	duration of second fixation on titles
2ndFixDur_Progress	-0.61**	duration of second fixation on progress indicator
2ndFixDur_Interaction	0.48*	duration of second fixation on interaction pane

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.0001$

In addition, Table 2 shows a number of fixation sequence features that correlate significantly with FIR. No significant correlations involved fixation on GRs alone, suggesting that successful learners spend more time interpreting graphics by relating them to other information. Looking between GRs or between GRs and hints was associated with higher error rates, while looking at GRs or hints in reference to the interaction pane was associated with lower error rates.

Table 2. Correlations between FIR (first-incorrect rate) and selected fixation sequence features

Feature	*r*	Feature Description
seq2_betweenGRs	-0.58*	2-point sequence between GR
seq2_GR-Interaction	0.54**	2-point sequence between GR and IP
seq2_Interaction-HintText	0.47*	2-point sequence between IP and HT
seq3_betweenGRs	-0.60**	3-point sequence between GR
seq3_Interaction-GR-Interaction	0.61**	3-point sequence between IP, GR, and IP again
seq3_GR-HintText-GR	-0.49*	3-point sequence between GR, HT, and GR again
seq3_HintText-GR-HintText	-0.47*	3-point sequence between HT, GR, and HT again
seq3_GR-Progress-GR	-0.44*	3-point sequence between GR, PI, and GR again

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.0001$
GR = Graphical Representations, IP = Interaction Pane, HT = Hint Text, PI = Progress Indicator

3.1 Prediction and Model Evaluation

While pretest scores alone had some predictive value with respect to FIR, gaze features provided a much more accurate model. A model containing both sets of predictors outperformed either set by itself. Pupil diameter was the least effective predictor of FIR. Top coefficients for the gaze-only model were consistently made up of a subset of the sequence features, the highest of which involved either interaction pane, GRs, or both. Following the correlations, the fixation sequence between the interaction pane, a GR, and back to the interaction pane was frequently the highest coefficient. Moreover, nearly all top coefficients involved GRs only when in relation to other AOIs. In general, more complex sequence features (3-point) were not more predictive than simple sequence features (2-point). The most accurate model contained only a binary indicator of which problem was given, which was likely representing difficulty. This model was not improved by the addition of gaze or pretest features. LGS was best predicted by pretest features alone, but pupil diameter did predict 5% above the majority class by itself. Table 3 provides a summary of these results.

Table 3. Average prediction accuracy for logistic regression models predicting FIR and LGS

Model	Accuracy (FIR)	Accuracy (LGS)
Majority Class	51.6%*	50%
Pretest Scores	58%	68%
Gaze Features	63%	53%
Pupil Diameter	51%	55%
Pretest Scores, Gaze Features	66%	68%
Problem	72%	44%

* Median splits of the criterion were not perfectly symmetrically distributed (Majority Class = High Error Rate)

4 Limitations & Future Work

While both gaze behavior and pupil diameter were predictive and informative, neither comprised the best predictors for our outcome variables. However, we expect these metrics to be more useful when problem difficulty and pretest scores are held approximately constant. While our simple gaze features were enough to outperform pretest-based prediction accuracy, not all of our results are directly interpretable in ways that can inform instructional design. Of particular interest is the finding that successful students relate graphical representations to other information as opposed to reviewing them in isolation. This spontaneous behavior may indicate that those students are using available resources more productively. This finding aligns with research on learning with multiple representations, which indicates that students need to integrate information presented across different representations [15]. Additional work will be needed to identify why exactly some of these features are such powerful predictors. While pupil diameter was somewhat predictive of learning gains, as a first glimpse at the utility of pupillometry in the wild, it was generally a very poor predictor of performance in our analysis. It is possible that even controlled ITS sessions introduce too much noise for pupil measurements to be useful. More likely, since cognitive phenomena detectable through pupil dilation are typically observed on the order of a few seconds, our dataset, which only contained average pupil dilation per problem, may very well have lacked the granularity necessary to capture these variations. In the time it takes to solve an entire, multi-step problem, learners may go through several positive and negative states of arousal, affect, cognitive load, and attentional shifts. Due to current difficulties with accurate time synchronization of the ITS and eye-tracking hardware, the current data set does not accurately represent smaller time scales. In future work, we plan to work with more granular, synchronized data to address this limitation.

5 Conclusion

Several of our analyses indicate that eye fixation and fixation sequence features are good predictors of how we have chosen to quantify student performance. While pupil diameter lacked similar predictive power, we expect that future experiments with higher temporal precision at smaller scales still holds considerable promise. Once a satisfactory set of features and level of granularity is established, one can explore the reasons why such features are indicative of the many fascinating aspects of the learning process and implement changes to instructional design based on this knowledge. Our results provide further motivation to explore the usefulness of eye tracking in educational research.

References

1. Poole, A. & Ball, L. J. (2006). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. In Ghaoui, Claude (Ed.). Encyclopedia of Human Computer Interaction. Idea Group
2. Tai, R. H., Loehr, J. F., and Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessment. International Journal of Research & Method in Education, 29(2), 185–208.
3. Tsai, M., Hou, H., Lai, M., Liu, W., and Yang, F. (2012), Visual attention for solving multiple-choice science problem: An eye-tracking analysis. Computers & Education, 58(1), 375-385.
4. Gomes, J. Yassine, M., Worsley, M., Blikstein, P. (2013) Analysing Engineering Expertise of High School Students Using Eye Tracking and Multimodal Learning Analytics. In Proceedings of the Educational Data Mining 2013 (EDM '13). Memphis, TN, USA. 375-377.
5. Wiedenbauer, G., & Jansen-Osmann, P. 2008. Manual training of mental rotation in children. Learning and instruction, 18, 1, 30-41. [12] WU, H. K., & SHAH, P. 2004. Exploring visuospatial thinking in chemistry learning. Science Education, 88, 3, 465-492.
6. Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, 91(2), 276.
7. Zekveld, A. A., Festen, J. M., and Kramer, S. E. (2013). Task difficulty differentially affects two measures of processing load: the pupil response during sentence processing and delayed cued recall of the sentences. J Speech Lang Hear Res.
8. Bornemann, B., Foth, M., Horn, J., Ries, J., Warmuth, E., Wartenburger, I., and Meer, E. (2010). Mathematical cognition: individual differences in resource allocation. ZDM, 42(6), 555–567.
9. Wierda, S. M., van Rijn, H., Taatgen, N. A., and Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proceedings of the National Academy of Sciences, 109(22), 8456–8460.

10. Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational Psychologist, 38, 1-4. doi:10.1207/S15326985EP3801_1
11. Martina A. Rau, Joseph E. Michaelis, Natalie Fay, Connection making between multiple graphical representations: A multi-methods approach for domain-specific grounding of an intelligent tutoring system for chemistry, Computers & Education, Volume 82, March 2015, Pages 460-485, ISSN 0360-1315, http://dx.doi.org/10.1016/j.compedu.2014.12.009.
12. Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. Journal of Experimental Psychology: Applied, 18(2), 178-191.
13. Mason, L., Pluchino, P., & Tornatora, M. C. (2013). Effects of picture labeling on science text processing and learning: evidence from eye movements. Reading Research Quarterly, 48(2), 199-214.
14. Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010) A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.
15. Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16(3), 183-198.