Exploring Pupillometry as a Method to Evaluate Reading Comprehension in VR-based Educational Comics

Kenya Sakamoto [*]
Osaka University

Shizuka Shirai[†]
Osaka University

Jason Orlosky[‡]
Osaka University
Augusta University

Hiroyuki Nagataki[§]
Osaka Electro-Communication University

Noriko Takemura[¶]
Osaka University

Mehrasa Alizadeh[‖]
Osaka University

Mayumi Ueda[**]
University of Marketing and Distribution Sciences / Osaka University

Abstract

Ascertaining the level of reading comprehension in a learner is often a challenging task. Although written tests and self-evaluations can provide feedback as to whether an individual understands a particular topic, they are not real time, do not necessarily provide a full picture of the reader's comprehension, and can be subjective.

In this paper, we present initial results of a study to determine better ways to evaluate a user's comprehension and understanding of educational comic books using pupillometry. Our system recreates the reading experience of an immunology comic book in virtual reality (VR), allows users to rate their comprehension of a particular section, and records eye data during the learning task. Through experiments, we explore the potential of this interface to facilitate learning and examine pupil metrics that might be used to automatically classify comprehension and understanding at the category (topic) level. We also discuss numerous design considerations that should be taken into account when designing future interfaces for evaluation of learning or comprehension.

Index Terms: Virtual reality, eye tracking, comics, education——

1 Introduction

Reading comprehension has been extensively studied in the past decades, with several major theories and models proposed [6] and disputes over appropriate approaches to teaching and researching reading [12]. Successful comprehension of a text is the result of a skillful coordination of many sub-processes, such as the control of eye movements, word recognition and parsing to name a few [3]. Eye movements particularly play a crucial role in reading as readers keep a sequence of fixations and saccades as they attempt to decode and comprehend texts.

In recent years, digital transformation has also had a large impact on education. Extended Reality (XR) technologies, for instance, are significantly changing the way we learn. XR devices equipped with various sensors for collecting biometric data, such as eye-tracking data, have also been recently developed, and it is therefore expected that learning systems utilizing XR technologies could provide tailor-made learning support based on learners' mental states such as subjective understanding or cognitive load.

In particular, augmented reality technology can add auxiliary information to the real world environment, and as a result, it could possibly facilitate learning in contexts where it is often difficult, if not impossible, to provide real-time, evidence-based support to learners by measuring and analyzing their cognitive processes. For example, it is difficult to provide real time analysis of the learning processes involved in reading and comprehending educational comic books since reading comprehension and diagnostic testing are typically carried out after the learning process. AR and VR combined with eye tracking may offer an alternative to conventional formative or summative assessments in both real and virtual learning spaces. Educational comic books have seen a remarkable growth in recent years and have begun to make their way into educational materials around the globe over the past decade. From science to art, to mechanics, comics have the potential to retain the attention of children and adults alike, and can be a fun way to learn new material. In fact, previous research has revealed that educational comic books increase learners' motivation [2, 5, 14]. If we can understand learners' mental states while reading educational comic books, we can improve their learning outcomes. However, classifying understanding at a category or topic level is no simple task. Eye metrics can often be noisy, and we do not yet know what metrics will correlate well with reading comprehension and understanding.

Moreover, within comic books, pages are often divided into panels using blocks and speech bubbles, making them structurally different from textbooks in several ways. Although there is research on the classification of individual word understanding [11], the difficulty of concentration-based tasks [9], and cognitive load when studying arithmetic [8], classification at the topic or category level has yet to be examined in depth.

The goal of this work is primarily to identify pupillometric metrics that correlate well with reading comprehension and understanding. Moreover, we are targeting users' abilities to grasp knowledge from a specific context. As such, we came up with the following hypotheses based on evidence from the literature:

- H1: Absolute pupil size will be greater for categories that are on average rated as more difficult to understand.
- H2: Changes (irregularities) in pupil size will be greater for categories that are on average rated as more difficult to understand.

To test our hypotheses, we conducted an in-depth experiment to reveal eye movements and changes in pupil size when learning immunology through comics. Results revealed that metrics such as self-reported level of comprehension and pre/post-test scoring were not well correlated to eye metrics.

More importantly, this experiment taught us that the evaluation of understanding or learning needs to be broken up into more specific types of learning. For example, memorization and logical understanding are both important for the process of learning, but each likely has a different method of evaluation and will probably produce different types of eye movements and pupil dilation.

[*] sakamoto.kenya [at] lab.ime.cmc.osaka-u.ac.jp
[†] shirai [at] ime.cmc.osaka-u.ac.jp
[‡] orlosky [at] lab.ime.cmc.osaka-u.ac.jp
[§] nagataki [at] osakac.ac.jp
[¶] takemura [at] ids.osaka-u.ac.jp
[‖] mehrasa [at] cmc.osaka-u.ac.jp
[**] mayumi_ueda [at] red.umds.ac.jp

2 Prior Work

Eye tracking has long been studied as a method for examining cognitive load and other cognitive metrics. Additionally, Marshall et al. have shown that pupillometry can help measure cognitive load for tasks occurring over several minute periods, further motivating our analysis of pupillometric measures [8]. Orlosky et al. found that eye movements, time spent on an item, and pupillometry can be indicative of word understanding [11]. However, since measures for other tasks like arithmetic or individual word learning are often specific to those tasks, we need to examine whether the existing features are also valid indicators for understanding comic books.

Several studies have been conducted regarding manga understanding. A key factor in comprehending manga is discovering referential links between text and graphics. In this regard, Rigaud et al. proposed a method for retrieving semantic associations between speech balloons (text) and comic characters (graphics) using geometric graph analysis and anchor point selection to enhance comics and manga understanding [13]. After testing their method with Japanese manga and European comic datasets, they found that in the presence of prerequisites such as anchors from balloons, comic character centroids, and tail positions, their approach is capable of detecting these associations with high accuracy. Other studies such as that by Cohn et al. examined whether comprehension of visual narratives such as comics are universal, and concluded that comprehension may be dependant on exposure to graphical systems. [1]. This also suggests that pupillometric measures may be contingent on prior experience with manga or comics.

Furthermore, it is important to clarify cognitive processes taking place during comic reading, given the lack of empirical research in this area. In order to bridge the gap, Laubrock et al. created corpora of eye movement recordings from a large number of participants and investigated their attentional selection by analyzing fixation locations and durations [7]. They observe that while initial fixations target visual elements, especially the main character, subsequent fixations are on text regions. Since cognitive processing of comics is distinct from mere text reading and image viewing [7], we set out to investigate learners' understanding of Japanese manga by measuring changes in their pupil size.

3 Methodology

This section summarizes the VR environment we built to simulate comics and conduct eye tracking, the experiment setup, and the methods used for recording and processing pupil data.

3.1 Simulation of Comic-based Educational Materials

To collect metrics with respect to each category, we built a VR environment that can replicate a typical comic book reading interface. To do this, we first created a plane at the same aspect ratio of the immunology comics in print and attached images of each page onto the plane in our virtual environment, as shown in Figure 1. This plane was placed directly in front of the user, as shown in Figure 2. To navigate backwards or forwards from each page, participants could click the left or right side of the pad on the controller.

Although the progression of the pages was almost exactly the same as the physical comics, we also included a duplicate copy of each page that contained a five point Likert scale for each panel or group of panels. In other words, the interface allowed participants to subjectively rate each block of content to obtain a metric of comprehension. When the controller was pointed at this scale, the pad buttons moved the rating from 1 to 5 rather than interacting with the manga page. The ratings from 1 to 5 corresponded to small billboards of text including: did not understand at all, did not understand well, understood to some degree, understood in general, and understood very well. This subjective rating would later be used in analysis as described below.

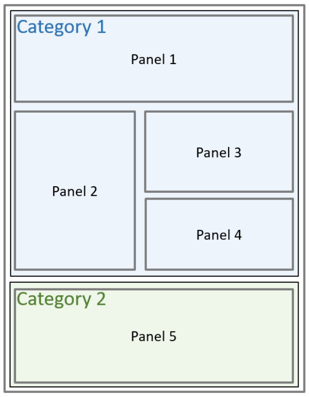

Figure 1: Image showing a typical comic page layout. In our study, we manually sorted each panel into one of 31 unique categories out of the two chapters. Some categories included multiple panels, while some others only had one panel.

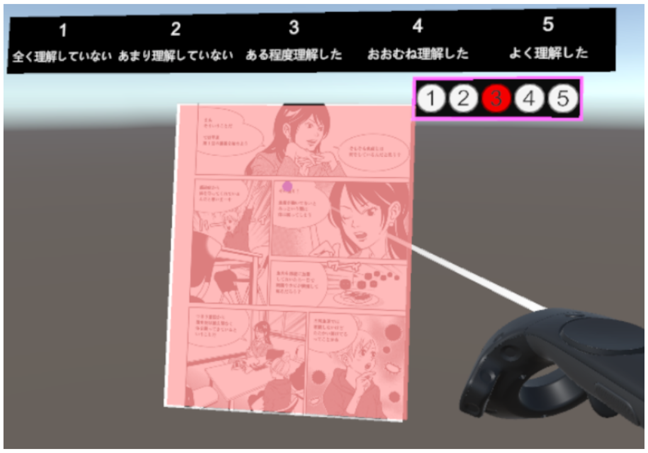

Figure 2: Image of the experiment interface showing a comic book image combined with a virtual Likert scale (top right). This allows participants to interact with the comic and then rate their perceived comprehension and understanding on the following page.

3.2 Experiment Design

The core goal behind this experiment was to establish a relationship between pupillometric metrics, the relative understanding of a category (group of panels), and resulting test scores. In order to conduct this experiment, we first had to pick a topic that the majority of participants would not be familiar with, and thus we opted for immunology as our topic of study. We used "The Manga Guide to Immunology," a popular resource for learning immunology [4].

Our ultimate goal is to determine the high level understanding of a specific category in real time, represented by a set of pages. By classifying understanding in this way, we should be able to identify both specific sections and topics that we can suggest the user to review prior to taking a test or evaluation. As such, we broke the comics into 31 different categories of content, including topics such as leukocytes, receptors, clones, and antigens. Each page was also evaluated individually to determine whether any tendencies existed for smaller areas of content. Any exits from the page area (i.e. looking away from the page or glancing at the controller), were excluded from analysis. Each category has from 1 to 9 panels, which include the frames that make up entire pages, as shown in Figure 1. As mentioned before, participants were presented with virtual, 5-point Likert scales with which to rate panels in these comics.

3.3 Participants and Experiment Setup

Using our comic reading interface, we set up an experiment with 11 participants (10 male, 1 female, aged 19 to 42, average 24.6, stdev. 6.3) to help us test our hypotheses. They majored in economics, law, physics, and engineering; however, we refrained from recruiting participants majoring in medicine or biology-related fields so that they would have as little knowledge of immunology as possible.

The hardware we used in our experiment was the HTC Vive Pro Eye. This VR system provides completely integrated eye tracking with relatively high accuracy at a reasonable price. Moreover the resolution of the display, 1440 by 1600 pixels per eye, was high enough that participants in our experiment were able to read the text without any major difficulties. The display was run on a GIGABYTE laptop with a i7-8850H processor and NVIDIA GeForce GTX 1070 GDDR5 8GB graphics card. Our software was built using Unity, version 2019.1.12f1.

3.4 Creation of Pre- and Post-tests

In order to evaluate participants' knowledge of immunology, we first created a pre-test to understand how much of the content they might already know. This pre-test contained questions specific to the text that evaluated the participants' general knowledge of the material. For instance, one question was "What is the concept of self-tolerance as it relates to immunology?" A similar test that evaluated the same topic areas was created as a post-test to determine whether test scores improved. We included 13 questions for the pre-test and 13 for the post-test, the questions of which were not completely identical, but contained essentially the same content.

We hypothesized that the answers to these questions would give us 1) knowledge of how well the learners absorbed the material during the study, and 2) allow us to correlate pupillometric measures to the increase in participants' post-test scores.

3.5 Procedures

The participants first sat down at a table and were explained that they would be using the new VR interface to learn about immunology and that their eye data would be recorded. After signing a consent form (approved by an IRB), they were first given a PC-based pre-test that evaluated their general knowledge of immunology. Once they had completed the test, they took the "Big 5" personality test [10] to give their memory some time to refresh between the pre-test and study phase. We used self-reported understanding using a five-point Likert scale as the ground truth for subjective understanding. After the personality test, the participants tried on the HMD and adjusted it so that they could see the content and read the text on the page. They were then told how to use the controllers to move backwards and forwards between pages. They proceeded to read and study two chapters from the immunology book and rated the content on each page using the duplicate page with Likert scales. Once the study phase (less than one hour for all participants) was completed, they were given a post-test that evaluated their post-study knowledge.

3.6 Data Recording and Filtering

In order to accurately determine the changes in pupil size, we had to eliminate significant noise from the eye tracking data. Upon scrutinizing the raw data, the Tobii tracking system resulted in fairly high variations in pupil size, sometimes over 10% on a frame to frame basis. This is of course physically impossible for the eye, meaning the noise was most likely coming from the tracking algorithm or camera. As such, we averaged values over 30 frames (approximately 300 milliseconds) to normalize the slow changes in pupil size. Blinks also interfered with accurate computation of pupil size, so we deleted the values provided by the eye tracker during blinks. To do so, we compared the median value of the last 30 frames with the average, and ignored any values with a difference of over 10%. Figure 3 shows pupil size data for about two seconds, where the orange line represents our recorded data, excluding points where the median value is 10% different than the average. This accounted for the vast majority of tracker noise and blinks throughout the experiment.

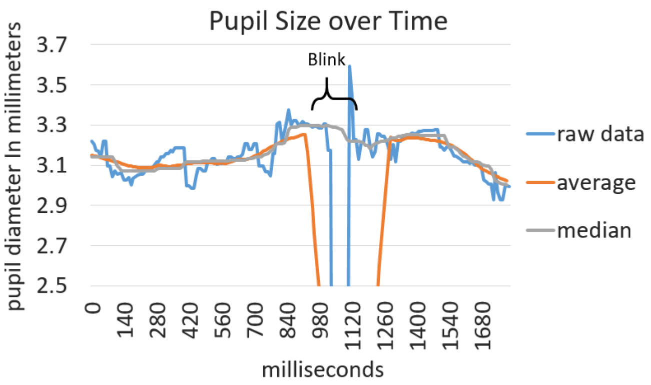

Figure 3: Graph showing the raw data, average, and median values of pupil size for approximately two seconds. Using these values, we can filter out both noise and erroneous pupil data during blinks.

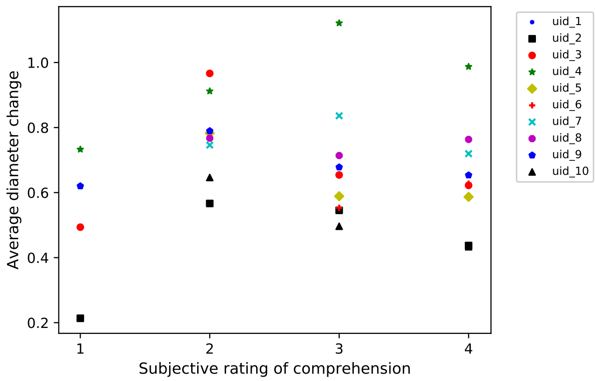

Figure 4: Plot of pupil size (vertical) versus self reported comprehension (horizontal) using the three categories of comprehension. No correlation between the metrics was found.

4 Results and Discussion

Our primary goal in this experiment was to try and associate self-reported comprehension and test score improvement with pupillometric measures. As such, in our quantitative analysis we first aggregated the subjective comprehension rankings of each category. Next, we ran a Spearman's rank-correlation using SPSS to determine if a correlation existed between increasing self-reported comprehension, pupil size, and changes in pupil size. We also ran post-hoc t-tests on pupil size with respect to self-reported comprehension for each difficulty level.

4.1 Correlation of Self-reported Comprehension with Pupil Metrics

As mentioned previously, Matsumoto et al. found that maximum pupil size increases with increasing difficulty of the task. Therefore, we first tried observing overall changes in pupil size according to rank, though the differences in size were not significant. We then ran a Spearman's rank-correlation to determine whether subjective score was correlated to pupil size. We had hoped that this would reveal certain tendencies on a per-participant basis, though no significant correlation was found. A plot showing the changes in pupil size

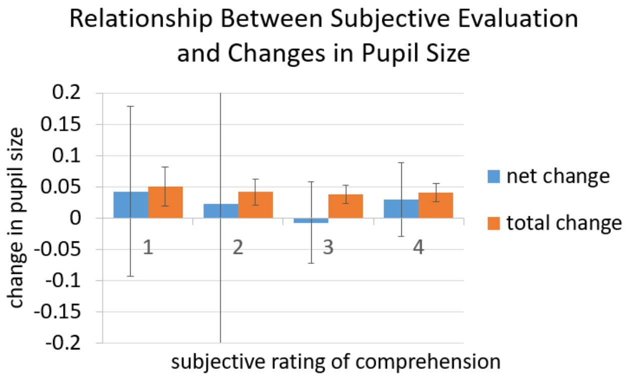

Figure 5: Graph showing average change in pupil size during a particular category according to user self reported ranking of comprehension.

according to subjective understanding is shown in Figure 4. Correlations ranged from almost -0.24 to 0.20, which indicated that pupil tendencies were essentially non-existent.

Figure 5 shows the average change in pupil dilation over time. This includes total change, which is the average magnitude (absolute value) of all changes over the course of the content for that rank, and net changes (average including both dilations and contractions) per rank. Again, pairwise t-tests with Bonferroni correction revealed no statistical difference between changes in pupil size versus rank.

The participants did improve their test scores between pre- and post- tests, with an average of 47% correct pre-test and 82% correct post-test. Though we also sought to correlate increases in test scores with pupil measures, the number of improvements made were not consistent and did not produce any observable tendencies. In other words, the differences in pre- and post- test scores were unevenly distributed amongst participants and questions, so we were unable to soundly analyze this data. Additionally, many questions covered content from multiple categories, so associating a particular gain in score with a particular category was difficult. As such, we have to reconsider the design of our questions.

4.2 Discussion

Upon examining the raw data taken from the eye tracker, it was imperative that we remove noise when computing changes in pupil dilation over time. If frame by frame differences were used to compute dilation, excessive noise would result in overly large changes in size. Moreover, utilizing a running median helped us remove pupil data that occurred during blinks, which would also have affected accuracy. We recommend that other researchers ensure data is properly filtered prior to analysis.

Secondly, the fact that none of our results turned out to be significant shows that analysis at the category level that was chosen by us may not be the best method for determining comprehension. Moreover, though we wanted to evaluate comprehension in general, the actual data that was evaluated was declarative in nature. Also the amount of the data for the “not understanding” rank category was low, which was likely prevented us from seeing a correlation. We plan to explore other types of analysis in future designs, for example at the page or panel/section level. Analysis at the bubble or call-out level may also be warranted for determining specific bits of information that might have been skipped over. We are also considering the use of the controller to allow the participant to specify particularly difficult areas when he or she feels confused or flustered. Overall, we need to concentrate on the specificity of information, meaning the specific time and location of the misunderstanding or lack of comprehension.

4.3 Realizations from Trying to Establish Ground Truth

In this experiment, the primary thing we learned was that the evaluation of understanding and comprehension are not only difficult to classify, but also difficult to define. For example, knowledge and understanding are divided into various facets such as logic, working memory, and the ability to infer information.

Moreover, the method for establishing a ground truth of understanding is also especially difficult. For example, self-ranked understanding via Likert scale might seem to be a reasonable method to obtain a ground truth, but in reality the participant himself or herself may not have a grasp on whether he or she understood the material.

A second difficulty is the association of test scores with particular areas of understanding. The general theory of testing is that when a learner has comprehensive knowledge of a particular field or subject, he or she will perform better on average on test questions. However, for the purposes of associating comprehension with specific eye movements, test questions do not necessarily correspond with specific points in the comic. This prevents us from using those test scores for establishing a good general pupillometric measure for classification. In other studies such as memory or mathematics tasks, the points at which the user does not understand a particular concept are clearly delineated, for example when viewing a new word. Conversely, when classifying subject categories in comics, the user is constantly trying to read and comprehend larger blocks of text over a greater time span with varying levels of difficulty.

5 Conclusion

In this paper, we present the results of an initial study testing different pupil metrics for evaluating the comprehension and understanding of educational comics. Through experiments and data analysis, we found that eye metrics were generally not well correlated to high level self-reported understanding or test scores in this context. More importantly, we deduced that the learning process likely needs to be divided into sub-categories of understanding such as memory or logic. Moreover, since ground truth for individual understanding is difficult to obtain, more focused experiments are necessary to find metrics that correlate with learner comprehension.

Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science, grant #19K21763 and the United States Office of Naval Research, grant #N62909-18-1-2036. We would also like to thank Prof Hiroshi Kawamoto of Kyoto University and Ohmsha, Ltd. for providing the comic materials with which we conducted experiments.

References

[1] N. Cohn. Your brain on comics: A cognitive model of visual narrative comprehension. *Topics in cognitive science*, 2019.

[2] I. Fuse and S. Okabe. Computer ethics education using video and manga teaching materials: learning effects and the order of using teaching materials. *Transactions of Japanese Society for Information and Systems in Education*, 27(4):327–336, 2010.

[3] P. B. Gough, W. A. Hoover, and C. L. Peterson. Some observations on a simple view of reacting. In C. Cornoldi and J. V. Oakhill, eds., *Reading comprehension difficulties: Processes and intervention*, pp. 25–38. Routledge, 1996.

[4] H. Kawamoto. *The Manga Guide to Immunology*. Ohmsha, 2014.

[5] T. Kogo and C. Kogo. The effects of comic-based presentation of instructional materials on comprehension and retention. *Japan Journal of Educational Technology*, 22(2):87–94, 1998.

[6] J. Kong. Theories of reading comprehension. In *Investigating the role of test methods in testing reading comprehension: A process-focused perspective*, pp. 9–29. Springer, 2019.

[7] J. Laubrock, S. Hohenstein, and M. Kümmerer. Attention to comics: Cognitive processing during the reading of graphic literature. In *Empirical Comics Research*, pp. 239–263. Routledge, 2018.

[8] S. P. Marshall. The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*, pp. 7–7. IEEE, 2002.

[9] A. Matsumoto, Y. Tange, A. Nakazawa, and T. Nishida. Estimation of task difficulty and habituation effect while visual manipulation using pupillary response. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, pp. 24–35. Springer, 2016.

[10] Y. Murakami and C. Murakami. Scale construction of a "big five" personality inventory. *The Japanese Journal of Personality*, 6(1):29–39, 1997.

[11] J. Orlosky, B. Huynh, and T. Hollerer. Using eye tracked virtual reality to classify understanding of vocabulary in recall tasks. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 66–667. IEEE, 2019.

[12] P. D. Pearson. The reading wars. *Educational Policy*, 18(1):216–252, 2004.

[13] C. Rigaud, N. Le Thanh, J.-C. Burie, J.-M. Ogier, M. Iwata, E. Imazu, and K. Kise. Speech balloon and speaker association for comics and manga understanding. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 351–355. IEEE, 2015.

[14] S. Shirai, H. Nagataki, I. Takenaka, Y. Takemoto, N. Tanabe, and S. Kanemune. Development and evaluation of course for learning how database works in information systems. 5(3):23–34, 2019.