Analysis of review quality by using gaze data during document review

Koki Saito (koki.saito@unisys.co.jp)
Nihon Unisys, Ltd., 1-1-1 Toyosu, Koto-ku, Tokyo, Japan
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

Shohei Hidaka (shhidaka@jaist.ac.jp)
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

Abstract

In software development, deliverables in an upstream process are reviewed to ensure their quality and to reduce error propagation to the downstream process. Methods are available for evaluating the review quality. In this study, we considered the defect detection process in a review of Requirement Definition Documents and tested a potential relationship between the gaze patterns and review quality. Specifically, we analyzed the relationship between the gaze patterns, with a primary focus on the blink rate, in a review of RDDs and detection accuracy. A significant nonlinear correlation between the blink rate and the detection accuracy was observed; moreover, the subsequent regression analysis also verified the blink rate as the best predictor of the review quality, notwithstanding the use of other gaze patterns. This result indicates that the blink rate is a major predictor of a type of review performance.

Keywords: gaze; blink rate; document review; review quality; signal detection theory; machine learning;

Introduction

In software development, it is important to ensure the quality of the specification and design document in the upstream process because they affect the quality of the deliverables in the downstream process. It is five to 200 times more expensive to correct defects in the downstream process than in the upstream process when we correct a low-quality deliverable affected by an ineffective specification (Boehm, 1981). Thus, it is preferable to maximize the defect detection in the upstream process.

In order to remove potential defects, it is common to review a document in an upstream process; moreover, numerous review methods have been used. However, individual differences in the review performance are likely to influence the review quality to a higher degree than the differences among the review methods (Uwano, Nakamura, Monden, & Matsumoto, 2007). A reviewer's performance also depends on the time limit for the task and the degree of the reviewer's concentration. Furthermore, although the defect detection rate based on the items indicated and the review rate are used for quantitatively evaluating the review quality, these indices by themselves are not adequate for accurately evaluating the review quality. First, the defect detection rate, for example, depends both on the quality of the reviewer and the quality of the document reviewed. As a result, we cannot assess whether a low detection rate implies low quality of the reviewer or high quality of the document. Second, these available indicators do not capture the different types of defects, such as simple typos, missing information, ambiguity, and misleading sentences. Accordingly, in this study, we explored a new indicator of the review quality, which characterizes the reviewer's performance and the potential types of defects. As a potential candidate for this indicator, we studied the gaze behavior in the document review task.

Recently, gaze data have been studied in software engineering (SE) to elucidate the cognitive process in various SE tasks such as code review (Sharafi, Shaffer, Sharif, & Gueheneuc, 2015). In SE, there are numerous studies targeting the review of a source code in the downstream process or review of "box and arrow" diagram such as Unified Modeling Language (UML). However, there are few studies on the review of documents in the upstream process (Sharafi, Guéhéneuc, & Soh, 2015). In fields other than SE, there has been studies on reading and understanding of narratives using gaze data (Augereau, Kunze, Fujiyoshi, & Kise, 2016; Campbell & Maglio, 2001; Okoso et al., 2015); however, there are few studies on the review process for detecting defects in a document.

Uwano et al., (2007) have defined the review process: "In the software review, a reviewer reads the document, understands the structure and/or functions of the system, then detects and fixes defects if any." They classified it into the three sub-processes: (1) reading, (2) understanding of the structure, and (3) detection/correction of defects.

Relevant to the three sub-processes above, past literature has reported the three major characteristics of eye blink as follows:

(A) An adult subject typically exhibits 20 eye blinks per minute (Bentivoglio et al., 1997).
(B) A task requiring certain external information such as reading tends to enhance external attention and reduce the number of eye blinks per unit time (Cho, Sheng, Chan, Lee, & Tam, 2000; Karson et al., 1981).
(C) A task requiring internal attention, such as mental arithmetic and association, increases the number of eye blinks per unit time (Cho et al., 2000; Karson et al., 1981).

In light of these observations, we hypothesize that the three sub-processes in the review process are related to the eye gaze patterns as follows: Process (1) is supposed to be associated with observation (B), wherein the rate of eye blinks would be reduced as it requires external information.

Processes (2) and (3) are supposed to be associated with observation (C), wherein the rate of eye blinks would be increased as it requires internal attention. Moreover, we suppose that both effective and ineffective reviewers are largely similar in the sub-process of reading (1); however, they would be different in the sub-processes of understanding (2) and detection (3). More specifically, we suppose that an effective reviewer would utilize more cognitive resources for the two sub-processes (2) and (3) than an ineffective one; as a result, a better reviewer would exhibit a higher rate of eye blinks per time.

Therefore, in this study, we executed an experiment that simulated a review process of a set of Requirement Definition Documents (RDD) and measured the reviewer's gaze patterns during the experiment. Then, we tested our above hypothesis by analyzing the relationship between the eye blinks and the review quality. In this experiment, we prepared a RDD material, to which we introduce defects; moreover, the review quality was defined based on whether these presumed defects were detected or not.

Our analysis of the review quality by using the gaze data revealed that the blinks were the most important component of the gaze data; a significant nonlinear correlation between the blink rate and the detection accuracy was apparent.

Experiment

In the experiment, each of the participants underwent two sessions: the review session and post-review session. In each trial in the review session, they were asked to review one page of the three types of the RDDs and then to mark the sentences with defects. After finishing 11 trials of the review session, they were asked to fill the demographic questionnaire.

Participants

We recruited 19 Japanese adults as the participants (16 male and three female) and the average age of them was 42.2 years (SD = 9.1), with nine of them in their 30s, four in their 40s, and six in their 50s. All of them were system engineer and nine of them had no RDD review experience. All of them had normal (corrected) vision.

Material

The set of original documents used in the review session were based on three types of RDDs that were in actual use at Nihon Unisys, Ltd. Each of the original RDDs was re-arranged such that each document had three pages of summary, three pages of functional requirement, three pages of non-functional requirement, and two additional sample documents—eleven pages in total. They were all in Japanese. On each page of a re-arranged RDD, we introduced a defect that was absent in the original document. In this study, the type of defects was the "omission" of certain necessary piece of information for requirement definition. A part of an original sentence was removed, which made the original definition ambiguous. In order to simulate a natural review process, we did not add more than a defect per page. As a result, we limited the number of sentences, including the one with a defect, to two per page; there were 17 sentences, including those with the defects, in the 11 pages. The demographic questionnaire included questions on age, gender, RDD review experience, document review experience, degree of concentration during review, and degree of comprehension to documents for review.

Procedure and Apparatus

In each trial, one page of the documents to be reviewed was presented on a computer screen; the participant's gaze patterns were measured by an eye tracker device during the document review. The eye tracker used in this experiment was gazepoint GP3HD eye tracker (Figure 1). The participants could spend as much time as they considered necessary for this review process.

After the review of each page, the participants were instructed to mark the sentences to be improved, on a printed document with the reviewed content; they were not informed about the type of defects introduced. This trial was repeated for 11 pages and the order of page was the same for all participants. The participants were not provided any break during the review trial. Moreover, they were instructed to maintain their head still as much as feasible in order to ensure accurate eye tracking.

After the review session, each of the participants was asked to answer the demographic questionnaire.

Figure 1: (left) Experimental situation, (right) eye tracker (set up at the bottom of the monitor)

Results

In the review session, the average, minimum, and maximum review times of the 19 subjects were 21, eight, and 40 min, respectively. In order to exclude the data with large numbers of eye tracking error, we performed the Smirnov–Grubbs test (Grubbs, 1969) to detect pages with valid fixation points less than 60% (Figure 2). Based on this test, we excluded data worth four pages (out of the total 209 pages—11 pages for each of the 19 subjects) from the rest of our analysis. We performed the subsequent analysis on the 205 pages of data with a sufficiently large rate of fixation.

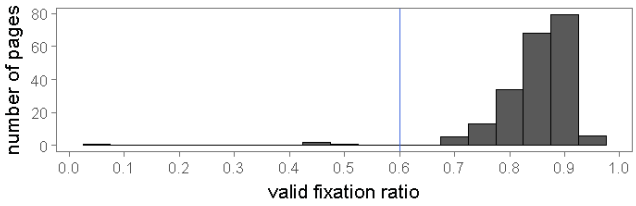

Figure 2: Histogram of valid fixation ratio each page

Review quality

In this study, we defined the correct review report for each unit of document based on the match between the participant's marked sentence and the sentence with a defect. Thus, the defect detection task was formulated as signal detection—the participant report a defect as either being detected or not, given a sentence with defect (signal in the ground truth) or otherwise (noise in the ground truth). We employed the signal detection theory (SDT) (Green & Swets, 1966) and treated the d-prime as an indicator of the accuracy of defect detection or the review quality. In the SDT, the respond bias and the sensitivity (d-prime) are distinguished from the rates of correct rating (the rate of defect detected marked to a sentence with defect) and false alarm (the rate of detect detected marked to a sentence without defect). The d-prime represents the deviation of the signal and noise distribution from the noise distribution as defined by

$$\text{d-prime} = \frac{M_{SN} - M_N}{\sigma_N}, \qquad (1)$$

where M_{SN} is the mean of the signal and noise distribution, M_N is the mean of the noise distribution, and σ_N is the standard deviation of the noise distribution. The d-prime is an indicator of the review quality; it can circumvent the effect of the potential response bias (the behavioral tendency to report detection regardless of the signal).

Analysis

In order to test our hypothesis, we analyzed the relationship between the blink rate and the review quality measured by the d-prime, in Analysis 1. In Analysis 2, we performed a model-based analysis of the relationship between the review quality and the gaze pattern not just the blink rate but also the other types of measurements such as fixation and saccade. The statistical analyses reported here were conducted with the free software R language (R version 3.4.1).

Analysis 1: Is the blink rate related to the review quality?

According to our hypothesis discussed in the introduction, the key sub-process in the review would require internal attention; thus, it would increase the blink rate. In order to verify this relationship between the blink rate and the review quality, the scatter plot of the blink rate and d-prime are shown in Figure 3. The corresponding correlation coefficient and other statistics are listed in Table 1. The maximal information coefficient (MIC) is a correlation coefficient calculated using mutual information; its application is feasible even with a nonlinear relationship. MIC-ρ^2 is an index of nonlinearity, and the maximal asymmetry score (MAS) is an index of non-monotonicity (Reshef et al., 2011). These results are summarized as follows:

- d-primes across the trials of the participants were distributed from - 1 to + 1.
- When d-prime was approximately zero, the blink rate was reduced from the mean blink rate and increased at non-zero d-prime values.
- The Pearson correlation coefficient between the d-prime and the blink rate was significant, although weakly negative.
- Both MIC and MIC-ρ^2 were large, and these together exhibited significant nonlinearity.

From the above facts, it was determined that the relationship between the blink rate and d-prime was a U-shaped or V-shaped nonlinear correlation, in which the blink rate was the smallest for d-primes near zero.

This result, both positive and negative d-primes across the trials, indicates the presence of two distinct groups of participants: One group detected the type of defects incorporated to a few sentences in the experimental manipulation; the other group detected the other type of defect (rather than only random sentences), which were not regulated explicitly in this experiment. Although we incorporated defects to a few sentences in the document, the original document (a RDD in use for some other purpose) is likely to have had certain other type of defects prior to this experimental manipulation. If so, the positive d-prime indicates the sensitivity to the expected type of defects incorporated in this experiment, whereas the negative d-prime indicates the sensitivity to certain unexpected type of defects originally in the RDDs.

Figure 4 shows the relationship between the d-prime and the response to noise in all the trials for each participant. We observed the general trend in these individual differences wherein those who exhibited a negative d-prime tended to detect "noise" as a "signal" (which may be interpreted as a defect by these participants) rather than the signal defined by the pre-experimental manipulation of the RDD. Thus, owing to the ambiguity of definition of the type of defects in the instruction, these participants were likely to detect the other types of defects (which were classified as "noise" by our definition).

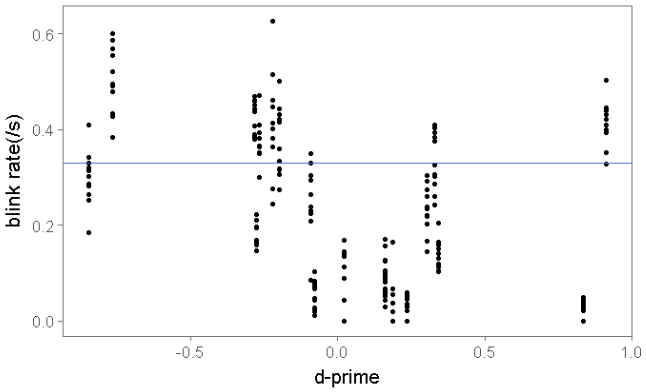

Figure 3: Scatter plot of d-prime and blink rate (blue line: mean blink rate in normal)

Table 1: Correlation coefficient between d-prime and blink rate

		Blink rate
d-prime	Pearson correlation	-.393
	p-value	.000
	MIC	.865
	MIC-ρ^2	.711
	MAS	.450

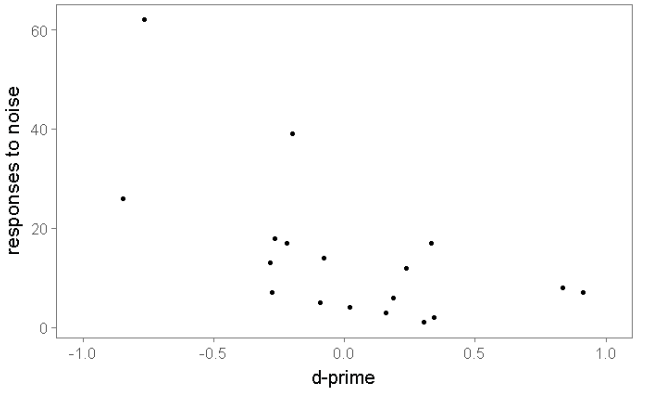

Figure 4: Scatter plot of d-prime and responses to noise

With respect to this interpretation of positive and negative d-primes, both positive and negative (non-zero) d-prime indicate higher sensitivity to certain types of defects; moreover, the blink rate was adequately correlated to the review quality of the potentially mixed types of defects.

This result appears to be evidence supporting for our hypothesis. However, it is likely that this result is caused by a spurious correlation owing to certain other features of gaze patterns, which are also correlated to the blink rate. Accordingly, in Analysis 2, we analyzed the d-primes, the indicator of review quality, with a collection of the other types of gaze features as well as the blink rate. Thereby, we evaluated the significance of the blink rate in the prediction of the d-primes, relative to the other types of gaze features such as fixation and saccade.

Analysis 2: Model-based analysis of review quality

In Analysis2, we constructed a model that predicts the type of detected defects measured by the positivity of the d-prime. Specifically, we employed a machine learning algorithm, random forest (RF) (Breiman, 2001) to predict the d-prime using the blink rate and other gaze patterns as the predictor. First, the set of features were calculated from the gaze patterns. Second, an RF regressor was constructed using the gaze features as a predictor of the d-primes in each trial. Then, we determined which gaze features is more informative for predicting the d-primes.

Extraction of features We extracted a set of 47 gaze features from the four fundamental gaze components: fixation, saccade, blink, and pupil (below). Forty six of these 47 features were originally defined by Bixler & D'Mello (2015); the blink rate was added to the list of features for the purpose of this study.

1. fixation: gazing on a single location
2. saccade: quick eye movement between fixation
3. blink: presence or absence of blink
4. pupil: size of the pupil

For each trial, the gaze pattern was characterized by these 47 features, and it was used to train the RF. As the RF calculated the importance of each feature simultaneously, the feature with low importance (with negligible significance for predicting the d-prime) could be removed.

We employed a sequential forward feature selection procedure using the RF as follows. First, the importance of each feature was calculated by RF using all the gaze patterns. Then, the root mean square error (RMSE) was calculated by RF using the highest importance feature. RMSE is the error from the actual value and is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}\,, \qquad (2)$$

where n is the number of pages, y_k is the specified d-prime of the kth page, and $\hat{y}_k$ is the d-prime predicted by the model. Second, RF was executed by adding the feature with the next highest importance, and the RMSE was similarly calculated. Third, the above process was repeated until the RMSE was the smallest, and the features effective for the review quality were extracted.

As a result, we obtained the 15 most important features extracted by RF, which are listed in Table 2. The blink rate was observed to be the most important. This indicates that the blink rate was the feature that was the most predictive of the d-prime.

Table 2: Features and their importance as extracted by RF

Rank	Features	Importance
1	blink rate	2.23
2	kurtosis of sccade duration	2.09
3	fixation duration / saccade duration ratio	1.93
4	number of blinks	1.93
5	max of saccade duration	1.88
6	range of saccade duration	1.80
7	kurtosis of pupil diameter	1.77
8	min of fixation duration	1.56
9	proportion of time spent blinking	1.50
10	mean of saccade duration	1.31
11	standard deviation of saccade duration	1.30
12	skew of saccade duration	1.27
13	proportion of horizontal saccade	1.14
14	median of saccade distance	1.08
15	median of fixation duration	1.01

Review quality prediction model Using the selected 15 features in Table 2, decision tree (DT), support vector regression (SVR), and multiple linear regression (MLR) models were constructed to predict the d-prime.

For validating these regression models, we performed a 10-fold cross validation (random split all trials) using their mean square errors (MAEs) and RMSEs. The MAE is defined as

$$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k|. \qquad (3)$$

The results calculated for each algorithm by constructing the review quality prediction model are listed in Table 3. SVM exhibited the lowest MAE and RMSE, whereas the RF exhibited the second lowest ones. To determine how effectively the model predicts the d-prime, we present the scatter plot of the d-prime of the data and the one predicted by SVM in Figure 5 and the corresponding correlation coefficient in Table 4.

To summarize the above results, 15 out of the 47 gaze features are significantly important for predicting the review quality measured by the d-prime. Among these significant features, the blink rate was observed to be the most important. This result of the model-based analysis is consistent with the observation in Analysis 1: The blink rate has a higher predictability than the other types of gaze features; thus, it is unlikely that the relationship between the blink rate and the review quality is the result of a spurious correlation.

Table 3: Review quality prediction model

		RF	DT	SVR	MLR
d-prime	MAE	0.224	0.323	**0.214**	0.283
	RMSE	0.304	0.451	**0.289**	0.361

Table 4: Correlation coefficient between actual and predicted d-prime

		d-prime (actual)
d-prime (predicted)	Pearson correlation	.750
	p-value	.000

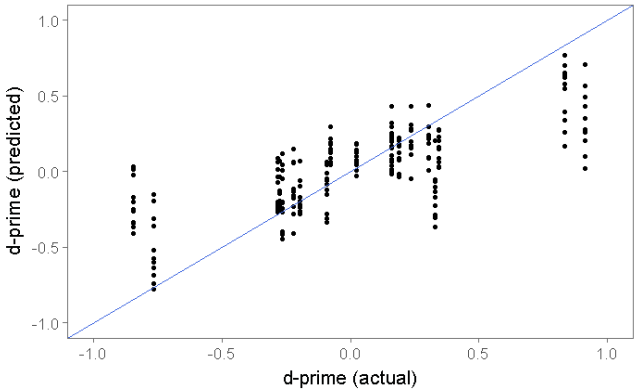

Figure 5: Scatter plot of actual d-prime and predicted d-prime

d-prime positive/negative classification model It is also intriguing whether we can classify the type of defects, which may be reflected as the positivity of the d-primes. Therefore, we next construct a classifier of the positivity of the d-prime by using the selected 15 gaze features. The algorithms adopted in this study were RF and support vector machine (SVM), which had exhibited a high prediction performance of d-prime in the experiment described in the previous section. Unlike the previous model-based analysis, we used the positivity of the d-prime as a class label rather than the d-prime value.

The classification accuracy is the coincidence rate between the predicted class and the specified class defined by

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \qquad (4)$$

where TP, TN, FP, and FN are the elements in the confusion matrix presented in Table 5.

Table 5: Confusion matrix

		Actual values	
		positive	negative
Predict values	positive	*TP*	*FP*
	negative	*FN*	*TN*

The classification accuracy for each algorithm is presented in Table 6. This result indicates that a classifier constructed upon the gaze features can predict the two potential types of defect detection with reasonably high accuracy 84%.

Table 6: Accuracy of d-prime positive/negative classification model

	RF	SVM
Accuracy	83.83%	**84.38%**

General Discussion

Blink rate and review quality

In prior study of SE, the gaze data has been used to elucidate cognitive process, however, the fixation and the saccade are often focused on and the blink rate is hardly taken consideration (Sharafi, Shaffer, et al., 2015). And it is also same trend in the study on reading and understanding of narratives (Augereau et al., 2016; Campbell & Maglio, 2001; Okoso et al., 2015). In this study, we focused the blink rate associated with each sub-process in the review and analyzed the relationship between the blink rate and the review quality.

In Analysis 1, we determined a nonlinear relationship between the blink rate and d-prime and that the blink rate was a U-shaped function of the d-prime estimated in each trial. This result is consistent with our hypothesis that the review quality (measured by d-prime) is related to the internal attention (measured by the blink rate). In Analysis 2, we tested the potential possibility that the relationship between the blink rate and d-prime is a spurious correlation owing to other confounding gaze features. We performed the regression analysis on the blink rate as well as the 46 other gaze patterns extracted from fixation, saccade, blink, and pupil. This analysis revealed that the blink rate was the most predictive of the d-prime; moreover, it indicated the blink rate to be a major gaze feature of the degree of review quality.

Limitations

It should be remarked that the result of Analysis 1, both positive and negative d-primes determined, was an indication of the likely presence of two potential groups of subjects detecting different types of defects owing to the ambiguity of the instruction for the review session. Considering this limitation of the experiment, it is feasible to have a few remarkable reviewers who detect both types of defects (the type defined and the other types not adequately defined in this study); such a reviewer may be evaluated near zero d-prime because he/she would detect both "signal" and "noise" according to our definition. Thus, in future works, an improved experimental design should have a list of defects covering most types of defects in the RDD material in order to prevent the problem of multiple types of defects.

Although we could not exhaustively classify all the types of defects using only the blink rate, Analysis 2 revealed that the positivity of the d-prime, indicating whether the detected defect was pre-defined or not, is classifiable with the blink rate and the other gaze features. There were numerous features on saccade duration in the 15 features. In general, these saccadic features capture gaze trajectory, and the saccade duration reveals the time of this trajectory. Thus, this result is likely to indicate the reading style such as reading order; moreover, the speed depends on the type of defects detected.

The set of RDDs used in this study was used for our customer's system development; its quality was supposed to be at least a specified level. However, it was likely that an immature RDD exhibited certain different types of potential defects than the defects introduced in this study. We cannot exclude the possibility that a reviewer's gaze pattern is affected by these mixed types of defects. This fact also necessitates a reconsideration of the experimental design that regulates the types of defects and investigates the relationship between the detection accuracy and the gaze patterns for each targeted defects.

References

Augereau, O., Kunze, K., Fujiyoshi, H., & Kise, K. (2016). Estimation of english skill with a mobile eye tracker. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16* (pp. 1777–1781).

Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., & Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Movement Disorders*, *12*(6), 1028–1034.

Bixler, R., & D'Mello, S. (2015). Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. In F. Ricci, K. Bontcheva, O. Conlan, & S. Lawless (Eds.), *User Modeling, Adaptation and Personalization* (pp. 31–43). Cham: Springer International Publishing.

Boehm, B. W. (1981). *Software Engineering Economics* (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Campbell, C. S., & Maglio, P. P. (2001). A Robust Algorithm for Reading Detection. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces* (pp. 1–7). New York, NY, USA: ACM.

Cho, P., Sheng, C., Chan, C., Lee, R., & Tam, J. (2000). Baseline blink rates and the effect of visual task difficulty and position of gaze. *Current Eye Research*, *20*, 64–70.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics. Signal detection theory and psychophysics*. Oxford, England: John Wiley.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21.

Karson, C. N., Berman, K. F., Donnelly, E. F., Mendelson, W. B., Kleinman, J. E., & Wyatt, R. J. (1981). Speaking, thinking, and blinking. *Psychiatry Research*, *5*(3), 243–246.

Okoso, A., Toyama, T., Kunze, K., Folz, J., Liwicki, M., & Kise, K. (2015). Towards Extraction of Subjective Reading Incomprehension: Analysis of Eye Gaze Features. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1325–1330). New York, NY, USA: ACM.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting Novel Associations in Large Data Sets. *Science*, *334*(6062), 1518 LP-1524.

Sharafi, Z., Guéhéneuc, Y.-G., & Soh, Z. (2015). A Systematic Literature Review on the Usage of Eye-tracking in Software Engineering. *Elsevier Journal of Software and Information Technology (IST)*.

Sharafi, Z., Shaffer, T., Sharif, B., & Gueheneuc, Y.-G. (2015). Eye-Tracking Metrics in Software Engineering. In *2015 Asia-Pacific Software Engineering Conference (APSEC)* (pp. 96–103).

Uwano, H., Nakamura, M., Monden, A., & Matsumoto, K. (2007). Exploiting Eye Movements for Evaluating Reviewer's Performance in Software Review. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *90*(10), 2290–2300.