Did you Understand this?

Leveraging Gaze Behavior to Assess Questionnaire Comprehension

Radiah Rivu
Bundeswehr University Munich
sheikh.rivu@unibw.de

Yasmeen Abdrabou
Bundeswehr University Munich
yasmeen.essam@unibw.de

Yomna Abdelrahman
Bundeswehr University Munich
yomna.abdelrahman@unibw.de

Ken Pfeuffer
Bundeswehr University Munich
ken.pfeuffer@unibw.de

Dagmar Kern
GESIS - Leibniz Institute for the Social Sciences
dagmar.kern@gesis.org

Cornelia Neuert
GESIS - Leibniz Institute for the Social Sciences
cornelia.neuert@gesis.org

Daniel Buschek
University of Bayreuth
daniel.buschek@uni-bayreuth.de

Florian Alt
Bundeswehr University Munich
florian.alt@unibw.de

ABSTRACT

We investigate how problems in understanding text – specifically a word or a sentence – while filling in questionnaires are reflected in gaze behaviour. To identify text comprehension problems, while filling a questionnaire, and their correlation with the gaze features, we collected data from 42 participant. In a follow-up study (N=30), we evoked comprehension problems and features they affect and quantified users' gaze behaviour. Our findings implies that comprehension problems could be reflected in a set of gaze features, namely, in the number of fixations, duration of fixations, and number of regressions. Our findings not only demonstrate the potential of eye tracking for assessing reading comprehension but also pave the way for researchers and designers to build novel questionnaire tools that instantly mitigate problems in reading comprehension.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**.

KEYWORDS

Reading Comprehension, Gaze Behaviour, Questionnaire

ACM Reference Format:
Radiah Rivu, Yasmeen Abdrabou, Yomna Abdelrahman, Ken Pfeuffer, Dagmar Kern, Cornelia Neuert, Daniel Buschek, and Florian Alt. 2021. Did you Understand this? : Leveraging Gaze Behavior to Assess Questionnaire Comprehension. In *ETRA '21: 2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Short Papers), May 25–27, 2021, Virtual Event, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3448018.3458018

ETRA '21 Short Papers, May 25–27, 2021, Virtual Event, Germany

ACM ISBN 978-1-4503-8345-5/21/05…$15.00
https://doi.org/10.1145/3448018.3458018

1 INTRODUCTION

A fundamental feature of human communication is comprehending ability, i.e the capability of a person to understand the information provided to them. We contribute to this trend by investigating eye movements while users experience text comprehension problems. Given the spread of ubiquitous computing and the usage of smart devices, the number of devices using eye trackers is increasing and will make it possible to detect reading problems on users' personal devices [12].

In this work, we focus on a particular reading and comprehension medium of questionnaires. Questionnaires act as a research instrument to gather data about people's beliefs, values, attitudes or behaviors [8] and allow for quick, statistical evaluation of the results [23]. Despite being fast and effective, questionnaires have several drawbacks. There are many inconspicuous factors, such as the order of the questions and the choice of the answer options, which influence quality [1]. One key aspect is to ensure that the data obtained through questionnaires is of high quality and leads to valid conclusions. Hence, respondents must be able to understand the questions to avoid distortion of the results.

To address this, creators of questionnaires usually do pre-tests before distribution [9]. However, such pre-tests are time-consuming, since participants need to be invited to the lab and interviewed. Despite much research in the area, avoiding questions' comprehension problems in questionnaires in a robust, timely, and unobtrusive way is still an open challenge. Recently, eye trackers have shown the potential to identify text comprehension [20]. Further, advances in miniaturization and mass production have continuously brought down the prices of these devices. With consumer-grade eye trackers readily available in the market for a few hundred dollars, measuring text comprehension at a larger scale becomes feasible.

In this paper, we attempt to approach this by addressing the following research questions: **RQ1:** Can text comprehension problems — both at sentence and at word level — be identified from gaze data as users read questions and answers? **RQ2:** Which features and combinations of features work best as indicators?

2 RELATED WORK

In this section, we discuss previous work on 1) detecting comprehension problems, and 2) how to mitigate them.

Previous works have explored a variety of approaches to detect comprehension problems. One of the directly linked cue to our reading behaviour is gaze behaviour [18]. Previous works have long explored how eye movement features can help uncover psychological states and recognize activities [22, 24]. Eye tracking is a powerful tool for understanding human attention as it can measure both the frequency of eye movements and the location of the gaze point [2, 3, 7, 10, 14]. McConkie et al. [15], present an analysis of gaze patterns to observe how users deal with visual distraction. This shows that gaze patterns can help us understand how the environment affects a reader. Prendinger et al. [19] measured user uncertainty and intention from gaze data. From the three presented approaches, two seek to detect uncertainty in different contexts: reading in a foreign language and trying to answer multiple choice questions. Fixations and regressions were found to be good predictors for detecting uncertainty. A similar study used gaze data to predict reading difficulties due to distraction [16]. The study reports that reading distractions and difficulties can be predicted with 80% accuracy and up to 15% better than by non-gaze features. We extend prior work by focusing on features and combinations of features which work best as indicators.

The analysis of gaze data allows for sophisticated interactive gaze-based applications which work to mitigate problems during reading. "The eyeBook" presented in [4] is an application for assisted and augmented reading. The system in this application tracks the current text being read by the user and provides apt effects such as illustrations and audio feedback. A similar approach by Sibert et al [21] assists students by providing automated responses. Researchers aiming to mitigate problems in questionnaires can benefit from our work by using features we identify as indicators.

Eye movements are an important part of visual attention for information intake activities, e.g., reading. They are primarily comprised of fixations (stationary phase) and saccades (rapid, ballistic eye movements phase). This paper, therefore, builds on the ability to use eye tracking as a method capturing gaze behaviour, to identify information comprehension. By combining this concept with machine learning techniques, we explore combining gaze features for comprehension problems detection. We hypothesize that by combining gaze features we can detect different comprehension problems. The gaze features that we take into account are *Perceptual Span, Fixations, Saccades and Regression and Pupil Dilation.*

3 PILOT STUDY

As a first step towards identifying problematic questions within a questionnaire based on gaze data, we analyzed data collected from previous work [17]. The data set was collected in a study in which eye tracking was incorporated in cognitive interviewing when pretesting survey questions. 83 people (39 males) aged between 18 to 76 (mean = 36) participated. Half (42 participants) answered the questionnaire while their gaze data was captured and the other half was the control group for a different research question that is not of interest to this paper. Hence, we focus our analysis on the 42 users who's eye data was collected.

The experiment was conducted as a repeated measure design where all participants were exposed to all conditions. The features to identify reading difficulties focus on: longer or repeated fixation of a word, repeated reading of a special word or passage, return from the answer to the question, a decision in the choice of answer, skipping the question. The data was saved in both pixels and millimeters notation. The questionnaire had 52 questions on various topics. We focused on 17 questions reported in the interviews and paper-based questionnaire as problematic ("vague sentences, undefined word, etc."). To record gaze data, a Tobii X120 Eye Tracker with 120 frames per second was used in combination with a 17-inch monitor, resolution of 1280×1024 and frequency of 120 Hz. For further processing of the data, Tobii studio software was used.

3.1 Results

The pilot study yields an important insight: gaze behavior strongly depends on whether participants experience difficulties with a specific word or an entire sentence. This allowed us to derive a set of features for each case (cf. Table 1 and 2). Data was analyzed using Support Vector Machine (SVM), NaiveBayes, and J48 in combination with ensemble methods. Features influencing the gaze data were derived using SVMattribute evaluation in WEKA algorithm.

4 STUDY: IDENTIFYING ISSUES ON A WORD AND SENTENCE LEVEL

We conducted a user study to investigate how different problems occurring in questionnaires influence gaze behavior.

4.1 Method

We designed a within subject repeated measures experiment in which participants had to answer an online questionnaire. This was different from the questionnaire used in the pilot study. 8 out of 24 questions were designed to cause reading difficulties (Table 7). During the experiment we recorded users' gaze path and collected qualitative feedback from semi-structured interviews. The experiment was conduced in a controlled lab setting with constant lighting conditions. Experiments took place on four consecutive days. To record gaze data, a Tobii X120[1] eye tracker was used. The screen was 24-inch with a recording rate of 120 frames per second.

4.2 Participants and Procedure

We recruited 30 participants (12 males), aged between 19 to 36. Participants were mostly students in different majors. Participants first signed a consent form and the purpose of the study was explained to them. Next, we asked them to sit on a chair, placed central to the screen and the eye tracker was calibrated to the participant. They then started filling in the questionnaire on their own. After the participant completed the questionnaire on the screen, they answered the same questions again on paper. This time they had to mark those questions where they had a problem in comprehension and describe the problem to the experimenter. They were asked to explain if the difficulty faced was the result of a particular word or an entire sentence. Where necessary, the questionnaire was re-discussed with the participant to eliminate ambiguity. If a

[1] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

Classification	Features/Characteristics	Description
Temporal features	Time required to answer the question, 1	Depends on the length of the question
3*Fixation features	Number of fixations per letter, 2	Number of fixations on the question, divided by the length of the question
	Fixation duration per letter, 3	Total duration of fixations on the question, divided by the length of the question
	Average duration of fixations, 4	Total duration of fixations divided by the number of fixations
4*Saccade features	Number of regressions, 5	Total number of regressions
	Number of regressions per question, 6	Number of regressions depending on the length of the question
	Average length of saccades per question, 7	Total length of saccades divided by the length of the question
	Number of returns per question, 8	Number of returns from the answers to the question
4*Pupil size features	Average pupil size, 9	Summation of pupil diameter per question divided by the number of samples taken
	Maximum speed of the pupil, 10	Maximum speed of the pupil movements
	Maximum pupil dilation, 11	Maximum dilation of the pupil size from the average size of a person.
	Number of high dilation I, 12	Number of dilation size of 2 * SD
	Number of high dilation II, 13	Number of dilation size of 3 * SD

Table 1: Features used for Detecting Reading Difficulties per Sentence

Classification	Features/Characteristics	Description
Temporal features	Gaze duration, 11	Summation of gaze duration on a sentence divided by the number of words
2*Fixation features	Number of fixations, 1	Total number of fixations on the sentence divided by the number of words
	Duration of fixations, 2	Total duration of fixations per sentence divided by the number of words
1*Saccade features	Number of regressions, 3	Total number of regressions per sentence divided by the number of words
8*Pupil features	Average pupil size within a word, 4	Summation of pupil diameter for the word divided by the number of samples
	Maximum pupil size, 5	Maximum pupil size within a word.
	Average speed of pupil dilation, 6	Summation of total pupil dilation speed divided by the number of samples taken
	Highest speed of pupil dilation, 7	Highest speed of pupil dilation per word
	Largest pupil size dilation, 8	Largest dilation of pupil size from the average pupil size per participant
	Number of high dilation, 9	Number of values within a word that dilated more than twice from the standard deviation of the average pupil size
	The reference word, 10	Which word had the largest pupil size was measured within a question.

Table 2: Features used for Detecting Reading Difficulties per Word

participant did not report any problems with the questionnaire, they were asked to explain their interpretation of the meaning of a word. There was no time-limit to finishing the study.

4.3 Results

Data was analyzed using Support Vector Machine (SVM), Naive Bayes, and J48. To increase classification quality, the mentioned algorithms were combined with ensemble methods. In particular, we used: *Bagging* splits the data set into multiple sets; [5, 6]; *Boosting* assigns weights to the objects in the data record and then selects data based on this weighting [13]; and *Stacking* combines several learning algorithms to improve predictive quality [13]. We used the following measures: accuracy "Acc" (comparison of the generated output with regard to the actual data); specificity "Spec" (true negative rate); sensitivity "Sens" (true positive rate); correction classification rate "CCR" (quality of the classification); chracteristics "Char"(the features used to determine prediction).

4.3.1 Evaluation on sentence level. Table 3 provides an overview of the accuracy achieved by using the different single features. In general, the number of regressions and fixation duration per letter are particularly well suited for prediction. Table 4 shows that an evaluation *across all participants* can achieve an accuracy of up to 74% on sentence level. If considering only 2 features, still an accuracy of up to 67% is feasible. When looking *at each person* individually,accuracy is slightly lower, both when considering all features, and when choosing a combination of only a few features.

Feature	Precision	Spec	Sens	SR
Number of regressions	0.73	0.99	0.13	0.75
Fixation duration per letter	0.70	0.96	0.24	0.77
Number of high dilations II	0.62	0.98	0.08	0.74
Number of fixations per letter	0.61	0.98	0.11	0.74
Average duration of fixations	0.60	0.97	0.11	0.74
Number of regressions per question	0.55	0.97	0.09	0.74
Average length of saccades per question	0.54	0.96	0.14	0.74
Time required to answer the question	0.50	0.99	0.05	0.73
Average duration of fixations	0.45	0.97	0.07	0.73
Number of high dilations I	0.35	0.98	0.03	0.73
Maximum pupil dilation	0.27	0.98	0.02	0.72
Maximum speed of the pupil	0	0.99	0	0.73
Average pupil size	0	0.98	0	0.72

Table 3: Precision of the different features for predicting difficulties on sentence level. Number of regressions and fixation duration per letter are most accurate. (SR=Success Rate)

4.3.2 Evaluation on word level. For the analysis on word level, we considered difference ranges, i.e. for how many pixels around a certain word we considered gaze behavior. On the y axis, we considered a tolerance of 10 px, on the x-axis (i.e. in reading direction) we considered 50 px as tn1, 100 px as tn2, and pupil size as tn3, respectively. Analysis was performed similarly as on sentence level. The results again show that both the number of fixations and regressions stand out clearly and thus form a good feature with 0.82 and accuracy of 0.76 in Table 5 and 0.81 and accuracy of 0.75 for per participant. Features related to pupil properties do not act as strong indicators when the subject had a problem with a word. However, the number of regression turned out to be quite predictive.

Relation	EM	Char	Acc	Spec	Sens	CCR
SVM:						
1:1	-	2,6,8,7,9,11	0.71	0.72	0.68	0.70
Total	-	1,2,3,5,7,8,10,11	0.65	0.95	0.28	0.79
1:1	bagging	2,8,7,9,11	0.69	0.70	0.67	0.64
1:1	boosting	2,6,8,7,9,11	0.71	0.72	0.68	0.70
1:1	stacking	1,2,3,5,13	0.66	0.61	0.76	0.68
NB:						
1:1	bagging	-	-	-	-	-
1:1	boosting	4,8,7,10,13	0.67	0.69	0.62	0.66
1:1	stacking	1,2,3,5,7,8,10,11	0.70	0.69	0.73	0.71
J48:						
1:1	bagging	1,2,3,5,7,8,10,11	0.74	0.74	0.72	0.73
1:1	boosting	1,2,3,5,13	0.64	0.57	0.77	0.67
1:1	stacking	6,10	0.67	0.66	0.69	0.68

Table 4: Evaluation across all participants on sentence level

4.3.3 Problematic Questions Analysis. Finally, the 8 questions supposed to cause a comprehension problem were individually examined. The classification was carried out on both sentence and word level. A consistently good classification quality was achieved. Only for two questions results are unsatisfactory. For one question this may be because of the fact that double negation require more thinking time. In addition, participants' behaviour during reading varies depending on the person, resulting in different characteristic values. This makes a prediction difficult, especially on word level.

To ensure that questions caused a reading difficulty, we mapped the number of participants who found this problematic to the question number (see second column of Table 6). Overall, a good predictive quality was achieved, with questions 16 and 18 being slightly worse. For question 16, this may be again due the double negation probably making participants think more about the meaning, hence creating a larger variance in time. Also, the cause of the problem is not confined to a single word but the whole sentence context is taken into consideration. This makes prediction difficult, in particular on word level. In question 18, the prediction on word level is not as successful as for the remaining issues. This may be because the problem of this question lies more in the answers than in the understanding. Giving preference to one of multiple answer options makes it potentially difficult to detect the problem at the word level.

5 DISCUSSION

We achieved a good prediction accuracy (0.7 in case of SVM) for detecting problematic questions. It was possible on both sentence and word level to reveal comprehension problems. The use of ensemble methods in general improves the results. In most cases, evaluation was better, independent of the person than for each person individually. However, both approaches can be considered good.

5.1 Reflection on Problematic Questions

To make sure that the collected gaze data reflect the difference between problematic and non-problematic question. We collected the qualitative thought of the participants regarding the question severity. We observe many participants found some questions to be problematic. All of the problematic designed questions were found problematic by the users except for one which included asking about two things in the same question.

Tol.	Relation	EM	Char	Acc	Spec	Sens	CCR
SVM:							
tn1	1:1	-	1	0.76	0.82	0.56	0.69
tn1	Total	-		1,3	1	0	0.94
tn1	1:1	bagging	1,3	0.76	0.81	0.59	0.7
tn1	1:1	boosting	1,2,3,4,5,11	0.75	0.82	0.55	0.69
tn1	1:1	stacking	1.3	0.69	0.68	0.71	0.6925
NB:							
tn1	1:1	bagging	1,3	0.70	0.85	0.36	0.60
tn1	1:1	boosting	1,2,3,10,11	0.72	0.83	0.43	0.63
tn1	1:1	stacking	1,3	0.67	0.66	0.70	0.68
J48:							
tn1	1:1	bagging	1,3	0.70	0.68	0.74	0.70
tn1	1:1	boosting	1,3	0.66	0.59	0.80	0.70
tn1	1:1	stacking	1,2,3,4,5,11	0.64	0.60	0.73	0.66
SVM:							
tn2	1:1	-	1,2	0.54	0.21	0.92	0.57
tn2	Total	-	1,3	1	1	0	0.94
tn2	1:1	bagging	1,2,3	0.67	0.62	0.78	0.70
tn2	1:1	boosting	1,2,3	0.66	0.57	0.85	0.71
tn2	1:1	stacking	1,2,3	0.66	0.62	0.76	0.69
NB:							
tn2	1:1	bagging	1,2,3	0.55	0.22	0.94	0.58
tn2	1:1	boosting	1,2,3,8,11	0.64	0.68	0.59	0.63
tn2	1:1	stacking	1,3	0.74	0.75	0.71	0.73
J48:							
tn2	1:1	bagging	1,3	0.74	0.76	0.70	0.73
tn2	1:1	boosting	1,2,3	0.62	0.58	0.71	0.64
tn2	1:1	stacking	1,3	0.75	0.76	0.71	0.74
SVM:							
tn3	1:1	-	1,2,3,5,11	0.79	0.83	0.66	0.74
tn3	Total	-	1,3	1	1	0	0.94
tn3	1:1	bagging	1,2,3,5,11	0.78	0.82	0.64	0.73
tn3	1:1	boosting	1,2,3,8,11	0.77	0.81	0.67	0.74
tn3	1:1	stacking	1,2,3,8,11	0.78	0.82	0.66	0.74
NB:							
tn3	1:1	bagging	1,2,3,4,5,11	0.83	0.91	0.45	0.68
tn3	1:1	boosting	1,3	0.76	0.80	0.64	0.72
tn3	1:1	stacking	1,3	0.71	0.74	0.63	0.69
J48:							
tn3	1:1	bagging	1,3	0.68	0.66	0.72	0.69
tn3	1:1	boosting	1,2,3,8,11	0.69	0.66	0.78	0.72
tn3	1:1	stacking	1,2,3,8,11	0.70	0.68	0.74	0.71

Table 5: Analysis for different tolerance ranges for all participants on word level

5.2 Assessment of the Individual Metrics

Number of fixations proved successful on both sentence and word level. On sentence level, the number of fixations depends on the length of the question. Thus, adapting it to the length of the question is necessary to be able to compare it with the others. This may also mean that the significance of this feature will be affected by the behaviour. On word level, it also depends on the length of the word. But since the word's prominence and its use in the context has an influence on the number of fixations, it is better not to depend on the length of the word. This will eliminate the feature from going down for long words and making it over-representative. In addition, it has been shown that the words around the problematic word should be taken into consideration as they also are affected and face an increased number of fixations. **Fixation duration** is related to the number of fixations, thus leading to similar results. To avoid faulty saccades, we included them, in contrast to prior

Question #	Problem	Perspective	EM	Sens	Acc
3	22	sentence level	bagging	0.95	0.78
			boosting	0.86	0.76
			stacking	0.95	0.78
		word level	bagging	0.77	0.73
			boosting	0.72	0.8
			stacking	0.81	0.81
5	12	sentence level	bagging	0.88	0.8
			boosting	0.88	0.8
			stacking	1	0.75
		word level	bagging	0.75	0.85
			boosting	0.875	0.63
			stacking	0.875	0.77
6	28	sentence level	bagging	0.89	0.80
			boosting	0.92	0.76
			stacking	0.85	0.8
		word level	bagging	0.85	0.92
			boosting	0.85	0.85
			stacking	0.92	0.78
11	23	sentence level	bagging	0.91	0.875
			boosting	0.95	0.84
			stacking	1	0.74
		word level	bagging	0.8	0.77
			boosting	0.71	0.78
			stacking	0.66	0.77
16	27	sentence level	bagging	0.67	1
			boosting	0.71	0.90
			stacking	0.78	0.84
		word level	bagging	0.57	0.61
			boosting	0.68	0.66
			stacking	0.68	0.72
18	9	sentence level	bagging	0.77	0.7
			boosting	0.88	0.8
			stacking	0.77	0.77
		word level	bagging	0.66	0.72
			boosting	0.75	0.64
			stacking	0.66	0.5
22	23	sentence level	bagging	0.86	0.8
			boosting	0.86	0.64
			stacking	0.82	0.79
		word level	bagging	0.81	1
			boosting	0.77	0.94
			stacking	0.86	0.82
24	29	sentence level	bagging	0.86	0.69
			boosting	0.82	0.68
			stacking	0.89	0.70
		word level	bagging	0.89	0.96
			boosting	0.89	1
			stacking	0.89	0.96

Table 6: Results for each problematic question.

work [11]. **Number of regression** achieved good results. Note, that this metric also depends on the question length. While on sentence level, this may reduce the power of this feature, on word level it is usually more pronounced. **Length of regression** can be an indication for a problem though, while reading fast, the reader makes wider recesses. If one uses the average of all regressions, this could be a feature, but it only shows a set of regular and short recesses, which will not enable solid conclusions. The number of returns from the answer options back to the question was not very clear in the evaluation, due to this feature being highly individual. Using the average **saccade length** as a metric has to be treated with care. For individuals, this feature can be a good metric, as it depends on personal reading style. The time required to answer a question depends on the person, as the answers were included in choices. A higher duration does not necessarily indicate a problem of understanding, but it may be due to a conscientious answer. **Pupil size** related features were weak. Reasons may include that many factors influence pupil size, including age. At the same time cognitive load weakly affects them. It is also hard to determine the reaction to a problematic word as there is always a delay for data processing. This makes it difficult to detect difficulties.

Two correlated features should not be used at the same time in a classification. This can cause individual properties to get more weight in the interpretation, leading to errors in the classification. Only the features "Number of fixations" and "Duration of fixations" should be considered together. Depending on personal reading behavior, a problem of understanding can be triggered by increasing the number of fixations and their duration, or both. To determine the fixations, both the velocity threshold method and the dispersion threshold method were used. Both largely coincide with the position of their calculated fixations, but the Velocity-threshold method resulted in overall more fixations. This influences the characteristics. However, since the determination of fixations is not clearly defined, we can not conclude which measure is better.

5.3 Evaluation of Feature Combinations

We found that combining several features the prediction quality can be significantly improved. Since the behaviour can vary based on the person, we encourage using feature combinations, e.g., average saccade length, number of returns per question, max pupil speed and max pupil dilation, based on different insights. As a result, individual behaviours can be better taken into account. The analysis revealed some combinations that frequently occur and achieve good results. On word level, these are number and duration of fixation, number of regressions and finally gaze duration.

6 CONCLUSION AND FUTURE WORK

We investigated how can gaze data can be leveraged to assess comprehension problem in questionnaires. We introduced 13 metrics for detecting problems on sentence level and 11 metrics for word level. Our user study showed that a good prediction of problematic questions is possible. Overall, it has been shown that prediction is possible on both sentence and word levels. The most useful features proved to be the number of fixations, duration of fixations and number of regressions due to high accuracy. However, weaker features in combination play an important role as well. With using combination of features, the predictive quality can be increased.

For future work we will use our insights to build a system capable of mitigating problems in real-time, for example, in the form of pop-ups or other means to assist the user.

REFERENCES

[1] Herman J Adèr. 2008. *Advising on research methods: A consultant's companion*. Johannes van Kessel Publishing.

[2] T Armstrong and BO Olatunji. 2009. What They See Is What You Get: Eye Tracking of Attention in the Anxiety Disorders. *Psychological Science Agenda* 23, 3 (2009).

[3] Othman Asiry, Haifeng Shen, and Paul Calder. 2015. Extending Attention Span of ADHD Children Through an Eye Tracker Directed Adaptive User Interface. In *Proceedings of the ASWEC 2015 24th Australasian Software Engineering Conference (ASWEC ' 15 Vol. II)*. ACM, New York, NY, USA, 149–152. https://doi.org/10.1145/2811681.2824997

[4] Ralf Biedert, Georg Buscher, and Andreas Dengel. 2010. The eyebook using eye tracking to enhance the reading experience. *Informatik-Spektrum* 33, 3 (2010), 272–281.

[5] Thomas G Dietterich. 1997. Machine-learning research. *AI magazine* 18, 4 (1997), 97.

[6] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

[7] Albert Hoang Duc, Paul Bays, and Masud Husain. 2008. Chapter 5.5 - Eye movements as a Probe of Attention. In *Using Eye Movements as an Experimental Probe of Brain Function*, Christopher Kennard and R. John Leigh (Eds.). Progress in Brain Research, Vol. 171. Elsevier, 403 – 411. https://doi.org/10.1016/S0079-6123(08)00659-6

[8] William Foddy and William H Foddy. 1994. *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge university press.

[9] Floyd J Fowler Jr. 2013. *Survey research methods*. Sage publications.

[10] Maite Frutos-Pascual and Begonya Garcia-Zapirain. 2015. Assessing Visual Attention Using Eye Tracking Sensors in Intelligent Cognitive Therapies Based on Serious Games. *Sensors* 15, 5 (2015), 11092–11117. https://doi.org/10.3390/s150511092

[11] Robin L Hill. [n. d.]. Roger PG van Gompel Martin H. Fischer Wayne S. Murray University of Dundee, UK. ([n. d.]).

[12] Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The Past, Present, and Future of Gaze-enabled Handheld Mobile Devices: Survey and Lessons Learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '18)*. ACM, New York, NY, USA, Article 38, 17 pages. https://doi.org/10.1145/3229434.3229452

[13] Ilias G Maglogiannis. 2007. *Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies*. Vol. 160. Ios Press.

[14] Matei Mancas, Vincent P Ferrera, Nicolas Riche, and John G Taylor. 2016. *From Human Attention to Computational Attention: A Multidisciplinary Approach*. Vol. 10. Springer.

[15] George W McConkie and David Zola. 1986. Eye movement techniques in studying differences among developing readers. *Center for the Study of Reading Technical Report; no. 377* (1986).

[16] Vidhya Navalpakkam, Justin Rao, and Malcolm Slaney. 2011. Using gaze patterns to study and predict reading struggles due to distraction. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1705–1710.

[17] Cornelia Eva Neuert and Timo Lenzner. 2016. Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology* 19, 5 (2016), 501–519. https://doi.org/10.1080/13645579.2015.1049448 arXiv:https://doi.org/10.1080/13645579.2015.1049448

[18] Alex Poole and Linden J Ball. 2006. Eye tracking in HCI and usability research. In *Encyclopedia of human computer interaction*. IGI Global, 211–219.

[19] Helmut Prendinger, Aulikki Hyrskykari, Minoru Nakayama, Howell Istance, Nikolaus Bee, and Yosiyuki Takahasi. 2009. Attentive interfaces for users with disabilities: eye gaze for intention and uncertainty estimation. *Universal Access in the Information Society* 8, 4 (2009), 339–354.

[20] Keith Rayner, Barbara J Juhasz, and Alexander Pollatsek. 2005. Eye movements during reading. *The science of reading: A handbook* (2005), 79–97.

[21] John L Sibert, Mehmet Gokturk, and Robert A Lavine. 2000. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*. ACM, 101–107.

[22] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 189 (Dec. 2018), 27 pages. https://doi.org/10.1145/3287067

[23] Linda A Suskie. 1992. Questionnaire Survey Research: What Works. Resources for Institutional Research, Number Six. (1992).

[24] Mélodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. 2012. Wearable eye tracking for mental health monitoring. *Computer Communications* 35, 11 (2012), 1306 – 1311. https://doi.org/10.1016/j.comcom.2011.11.002

A APPENDIX

Category	Example
Q3: Made-up technical term; not relevant for understanding question	The free trade agreement TTIP would force many companies to resummate. How does this affect your view on the economic development in Germany?
Q5: Technical term from medicine	How would you find it if an insurance company's contribution amount is defined based on your personal history?
Q6: Technical term from banking	What do you think if your bank sends a remittance advice to secure bank transfers?
Q11: Using similar word with a different meaning	On the warnings of many cigarette boxes is the saying Smoking seriously harms you and others. In your opinion, does this apply?
Q16: Using a double negation	Isn't it true that we do not accept bad manners?
Q18: Asking two things in the same question	How do you like the German rock and Folk music?
Q22: Altering a well-known phrase	The manslaughter penalty is usually between five and 15 years in Germany. Not everyone considers this punishment resistant. How do you feel about it?
Q24: Made-up term; influencing meaning of sentence	What do you think about the increased use of media that causes many people to repatriate more often?

Table 7: Problem Category Examples