Factors influencing trust in Wikipedia: An eye-tracking approach.

Malte Risto

Master Thesis

Faculty of behavioral sciences

Enschede, 31 January 2010

Subject:	Masterthesis C&M
Subject code:	293496
First Supervisor's:	Dr. Matthijs Noordzij
	Teun Lucassen, MSc
Second Supervisor:	Prof. Dr. Jan Maarten Schraagen

UNIVERSITY OF TWENTE.

Contents

Abstract

The goal of this study was to gain insights into the overt visual attention behavior of users evaluating the credibility of Wikipedia articles and to compare it to verbal reports of important credibility cues. Furthermore it was investigated how fixation distributions are affected by the credibility assessment and a time constraint, compared to a control condition. A total of forty-three university students read six Wikipedia articles for general information while half of them also had to rate the articles for their credibility. Both tasks had to be carried out under a time constraint of two or five minutes. Eye tracking was used to obtain fixation frequencies on distinct article element categories (e.g. Introduction, Pictures or References). Comparison of fixation distributions showed significant frequency differences between element categories as well as between top, middle and bottom parts of text blocks. Partly these results seem to be in stride with verbal reports. No significant influence of task or time constraint was found. Furthermore subjects credibility ratings reflected quality differences between articles but again showed no effect of time constraint. The results lead to the conclusion that different article elements are approached with different viewing strategies and that verbal reports can help interpreting fixation distributions.

1. Introduction

1.1 Background

Within the last years collaborative information repositories (CIR) have had a large impact on the way we obtain information from the internet. The term CIR was coined by McGuiness and colleagues (2006) to describe web based content management systems whose contents can be read but also written by its users. A famous implementation of this technology is the wiki[1] which can act as a feature, or even build the structural basis for websites. Wiki's are not bound to certain topics, rather they are platforms to organize the process of information exchange among their users, where information is usually structured and presented in the form of interlinked articles.

Founded nearly a decade ago Wikipedia is one of the best known and most accessed wikis[2]. It has established as a multilingual online encyclopedia whereby the English version counts over three million articles[3]. It can be argued that the success of Wikipedia partly stems from its reliance on the wiki technologies. The underlying thought is that everyone with an internet connection and a web browser should be able to join the community and participate in creating and editing the knowledge database. To this point its accessible structure and openness for modification stimulated the generation and modification of articles adding to the bandwidth of covered topics.

However critics of Wikipedia often point at topics that are connected to the sites functionality as a wiki. Denning and colleagues (2005) published a list of risk factors for Wikipedia that can be related to its collaborative, user-generated nature. According to the authors risk lies in the uncertainty about the accuracy of the content, the motives and levels of expertise of the authors, the stability of articles, the coverage of the topic, and the kind of sources that are cited.

---

[1] From the Hawaiian term meaning: fast.
[2] According to website usage statistics Wikipedia resides among the ten most visited sites on the internet (Alexa Internet, Inc., 2010).
[3] In comparison the online version of the Encyclopædia Britannica contains about one hundred twenty thousand articles (Wikipedia:Size_comparisons, 2010).

The lack of control and reliability has deemed Wikipedia as an undesirable research tool among librarians and scientists (Chen, 2007). Yet Wikipedia is being cited in peer-reviewed journals ever more often. Table 1.1 shows search results produced when searching the ScienceDirect (www.sciencedirect.com) database for articles that refer to Wikipedia, without containing Wikipedia in their abstract, title or keywords.

Table 1.1
Articles referencing Wikipedia without discussing Wikipedia

Year	Frequency
2003	1
2004	9
2005	30
2006	129
2007	319
2008	443
2009	663

Note. Retrieved from ScienceDirect database. http://www.sciencedirect.com/

Wikipedia relies on its community in order to fact check and correct false entries. Several studies have tested the sites self-healing abilities as researchers deliberately inserted errors and measured the time these errors were corrected by the community (Halavais, 2004; Magnus, 2008). Error correction times were fairly low; most of the errors were corrected within the first three to twenty-four hours after insertion.

In 2005 Giles asked if Wikipedia can be considered a reliable source of information in comparison to an online encyclopedia maintained by a privately held company. Academic reviewers that were considered experts in their disciplines compared the accuracy (e.g. factual errors, critical omissions and misleading statements) of Wikipedia articles to those with matching topics taken from the online version of the Encyclopedia Britannica. From 42 randomly selected “general science” articles reviewers found 162 mistakes in Wikipedia

versus 123 in Britannica. Although this study suggests that Wikipedia shows error rates comparable to other encyclopedias, it is far from being error free.

A study by Lim (2009) found that among students a widespread reason to use Wikipedia is the need for background information about lesser known topics - especially factual information. The majority of students reported using Wikipedia for non-academic personal purposes. Furthermore the author notes that students tend to have positive past experiences with Wikipedia, yet lack a comparable positive perception of Wikipedia's general information quality. This poses the question if students actively asses the information quality of Wikipedia articles to decide whether or not to use its information. And if so, which information on the site do they use to evaluate their trust in the information?

Finding a general definition of trust seems to be a problem in the scientific literature. As a result Philosophers, Psychologists, Managers and Marketers developed definitions that best suited their fields of interest (Wang & Emurian, 2005). The following, psychological definition by Rotter (1967) denotes trust as "[…] an expectancy held by individuals or groups that the word, promise, verbal, or written statement of another can be relied on." (p. 651). This definition also fits well in the context of trust in information provided by an online encyclopedia.

As we look closer, the words trust and credibility are sometimes used interchangeable to describe the same concept. According to Fogg and Tseng (1999) this inconsistent and imprecise use of the words poses a semantic problem to anybody studying these concepts. Trusting is an activity carried out by a trusting person (or trustor). For the development of trust the trustor needs to judge the reliability and dependability (Fogg & Tseng, 1999) of a trust receiving object (or trustee). Also part of this judgmental process is the attribution of credibility a concept that is related to the believability of (information provided by) a trustee.

According to the literature (Fogg & Tseng, 1999; Metzger, 2007) credibility again has two dimensions: expertise and trustworthiness.

Fogg and colleagues (2001) proposed that credibility is not a quality inherent in objects, persons or information but can only be perceived by the observer. Accordingly, credibility related information that is not perceived will not influence the credibility evaluation. For example only the perceived elements of a Wikipedia article will contribute to the final credibility rating of that article. Furthermore the evaluation of credibility is considered an iterative process (Hilligoss & Rieh,2007). Perceived credibility is the product of an assessment which in turn consists of several credibility judgments. In sum, credibility is a perceived quality that relates to the believability of information. Also, consistent credibility judgments should lead to enhanced reliability and dependability ratings, both sub constructs of trust. On this background, when studying trust in Wikipedia it seems interesting to look at the processes that underlie credibility assessments of Wikipedia articles.

Rather than being deterministic the relationship between information (e.g. website content) and credibility is influenced by situational and personality factors (Corritore et al. 2003). The literature proposes several models of credibility evaluation in online environments that vary with respect to their level of abstraction and applicability to different sorts of websites. Some have been designed with specific websites in mind others were designed to be applicable to many different sorts of websites. The way Wikipedia and wikis in general congregate and distribute information over the internet and the underlying collaborative nature can be of significant influence on the perceived trustworthiness of users. To this point it has not been studied whether any of the existing models can be successfully applied to Wikipedia.

Wathen and Burkell (2002) propose an iterative model of how users judge the credibility of online information. In their model the credibility assessment is divided into

three stages of user website interaction. The first thing a person will notice when entering a website is the sites direct visual appearance and presentation (e.g. colors, graphics, typography), its usability and interface design and the general organization of information. At this stage a judgment about surface credibility is made. At the second stage a more in depth evaluation is made about the sites message (e.g. content, relevance, currency) as well as its source (e.g. expertise, trustworthiness). The first two stages solely deal with the sites content and its presentation. At a third stage the interaction of the sites content with the users cognitive state is assessed. At this point external factors such as the need for information, the stressfulness of the situation or the prior knowledge can have influence on the processing of the perceived information. For example, a strong need for information may lead to a weaker impact of surface credibility on the overall evaluation as the visitors focus is more on the message and the source.

The proposed process seems straightforward and the factors identified may be valid predictors of the outcome of a credibility assessment. However there has been a lack of empirical support that users actually behave in accordance with the model (Wathen & Burkell, 2002; Metzger, 2007). The model provides a comprehensive list of factors that can influence the credibility assessment however little information is given on the particular role of each factor and how it is affecting the credibility rating.

Metzger (2007) argued that only a few of the factors previously deemed important for a credibility assessment were actually evaluated by subjects. For example Scholz and Crane (1998) found that students often base their credibility evaluations on only one or two criteria. Flanagin and Metzger (2000) also found that checking online information against criteria previously identified as credibility related (i.e. accuracy, authority, objectivity, currency, and Coverage) occurred rarely to occasionally under college students as well as general adult internet users. There seems to be a discrepancy in what people deem important for the proper

evaluation of credibility and the actual criteria being evaluated. In this Study eye tracking is combined with questionnaires to shed light on the relationship between attended elements of a Wikipedia article and verbal reports of their importance in a credibility assessment. Cases are discussed where the findings differentiate. Furthermore it is investigated how visual attention processes in a credibility assessment are affected by external factors. This study is the first to look at general visual attention distributions of Wikipedia users in a credibility assessment. Findings should give insights in the actual process of credibility evaluation that underlie a certain credibility rating.

1.2 Eye tracking

Lucassen and Schraagen (n.d.) used think aloud protocols to identify article elements that were attended during credibility assessment. The study found that textual elements (esp. comprehensiveness, correctness and length), references (esp. amount, quality) and pictures (esp. quality, relevance) were the three most mentioned article features in a credibility assessment.

Think aloud protocols are very well suited to uncover users cognitive processes while carrying out a task. On the other hand the method is rather obtrusive and can lead to more conscious evaluation behavior than it would occur naturally. It can affect the subject's temporal distribution of attention as people spend more time attending to certain site elements while trying to translate their cognitive processes into language. Tasks may become unnatural because people are not used to speak out every thought they have (Nisbett & Wilson, 1977). Eye tracking is used as a more unobtrusive way to study attention processes on websites. Like think aloud protocols eye tracking can denote attended site elements in a credibility evaluation task while interfering to a lesser extent with the user's natural behavior. Therefore both methods may complement each other.

Eye tracking describes the process of recording the ocular movement of a person also called gaze movement. To provide a common ground for research the literature agreed on a definition of certain gaze movements. One of the most studied gaze features, the fixation is defined as moments of nearly motionless gaze in a certain area over a certain amount of time (Rayner, 1998). No universal agreement exists about the exact time span and size of the area. Among other things these measures depend on the task for which eye tracking is used. Many researchers handle intervals between 100 – 300 milliseconds (Pan et al. 2004, Duchowski, 2002, Beymer et al., 2007).

Fixations gain importance under the assumption that visual attention is aimed at a specific area in the visual field when this area is fixated. This has been called the "eye-mind assumption" (Just & Carpenter, 1980; Rainer, 1998). In this sense eye tracking can provide qualitative and quantitative data of the distribution of overt visual attention. However the gaze position, even a fixation can give no guarantee that the viewed item has been processed to the point of recognition or access to working memory (Duchowski, 2002). This phenomenon is further documented in studies on inattentional blindness (Simons, 2000) and change blindness (Simons & Levin, 1997).

According to Viviani (1990) fixations are associated with the cognitive processing of visual information. In past research the duration and amount of fixations on certain areas of interest (AOI) has been of special interest. Fitts, Jones and Milton (1950) studied fixation frequency and duration of Aircraft pilots during landing approaches. They proposed that the amount of fixations in an area indicated the degree of importance of that area while the study compared the distribution of attention on visual flight displays. Wikipedia articles as well are complex compositions of different visual forms of presentation (e.g. text, lists and pictures).Yet it is assumed that different ways of presenting information on Wikipedia evoke different eye movements. For example reading text in general may yield different fixation

patterns than looking at pictures or references. Although the distribution of visual fixation can denote the importance of article elements, it may be misleading just to compare the fixations on text with fixations on pictures and conclude that text is more important. Element wise analysis of changes in mean fixations in reaction to external factors seems a promising method to study these effects without relying too much on the comparison of different article element categories.

1.3 Hypotheses

As noted earlier a think aloud study by Lucassen and Schraagen (n.d.) identified textual elements, references and pictures to be the most mentioned element categories when rating the trustworthiness of Wikipedia articles. According to the different actions associated with different categories we would expect the majority of fixation on textual elements followed by pictures. With respect to references, fixation frequencies might depend on whether the subject is interested in the amount or quality of references. Quality assessment would call for a more elaborate examination resulting in higher fixation frequencies. A more heuristic estimation of the number of frequencies would yield to lower fixation frequencies.

H1 - Article element categories will show different fixation frequencies suggesting distinct visual approach strategies.

The assumption that people always engage in an elaborate credibility assessment had been questioned in the past (Metzger, 2007). Following this claim it may be asked whether credibility evaluation can be considered a natural behavior while reading Wikipedia articles. If so fixation frequencies on different element categories would show differences for just reading an article compared to reading an article while rating its credibility.

H2 - The task of assessing the credibility of an article will have an effect on the distribution of visual attention in that article.

According to Wathen and Burkell (2002) situational factors can influence the importance of trust related elements. Stress in the form of time pressure can influence a final trust rating and may already have an effect on the distribution of visual attention in an article. Metzger (2007) proposed a dual processing model of credibility evaluation[4]. She argues that differences in motivation and ability trigger different evaluation processes. The model differentiates between systematic and heuristic evaluation as the two modes of information processing. Basically, if the motivation as well as the ability to process a message are high, people are expected to engage in a more effortful, conscious and systematic evaluation of the augments that will result in a strong attitude towards the message. If one of the factors is low people are expected to express a less effortful, less conscious and more heuristic processing of peripheral message cues. In this study ability may be influenced by a time constraint. It is assumed that, given less time a subject would rely less on direct examination of the text and more on peripheral cues for a credibility assessment.

H3 - Time pressure will lead to less fixations on text elements in favor of article references for subjects in the credibility assessment condition.

Wathen and Burkell (2002) proposed that evaluations of surface and message credibility contribute to the final credibility rating. Wikipedia handles an internal rating system where article quality is evaluated based on predefined criteria (see Methods for further explanation). Note that the primary goal of this study is to explore the effects of external factors on the process of a credibility assessment rather than the product. Therefore we are interested how articles of differing quality are perceived under differing time constraints. We assume that with more time subjects will be able to make better assessments of the factors that influence their rating to the positive or negative. Therefore time may change the amplitude of the final rating.

[4] For another popular example of dual processing models see Petty & Caciappo (1981).

H4 - Differences in article quality will be reflected in subjects credibility ratings.

H5 - Time will amplify the effect that quality has on credibility ratings.

As mentioned earlier eye tracking can lead to different conclusions when compared to verbal reports (e.g. think aloud studies). Aside from fixation distributions we are therefore also interested in the factors deemed relevant in a credibility assessment as these factors can support the interpretation of those distributions. For example providing cues whether frequently viewed areas are really seen as important sources of credibility related information.

2. Method

2.1 Participants

A total of 43 college students[5] (32 female) aged between 18 and 25 (M: 20,55; SD: 1,81) completed the experimental procedure. Of those 17 participants had the Dutch nationality 26 were German. Contact lenses and glasses were exclusion criteria as they could interfere with the eye tracking procedure. Also subjects with dyslexia were excluded as their reading disorder could influence fixation frequencies in the text. All participants gave their informed consent for inclusion in this study and received credit points for their participation.

2.2 Apparatus

Gaze data was recorded simultaneously for both eyes at a sample frequency of 60Hz on a video based eye tracker that was placed under the monitor. Figure 2.1 shows the experimental setup of the two computers used for gaze tracking, stimulus presentation and data storage. The first computer was connected to a set of cameras that recorded head movements and corneal reflections and processed this data using FaceLab version 4.5 (Seeing Machines Inc., Acton, USA). The recorded gaze coordinate data was sent via local area network to the second computer running GazeTracker version 7 for FaceLab

[5] Mostly first year psychology students.

(Eye Response Technologies Inc., Virginia, USA). GazeTracker was used to present the stimulus articles and saved the gaze coordinates. Stimulus articles were pre-saved locally and presented in Internet Explorer 7 (Microsoft Corporation, Redmond, USA).

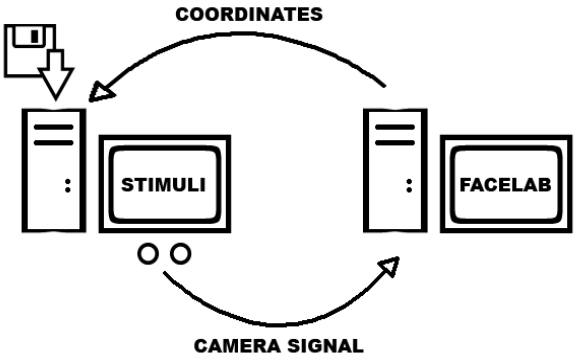

Figure 2.1 PC setup used in the experiment.

To ensure that eye fixation coordinates recorded by FaceLab were correctly adjusted to account for stimulus shifting events (e.g. scrolling) added by GazeTracker, computer times were synchronized using the Network Time Protocol (Mills, 1991). Offset between both computer clocks, measured before every test session, was always below the required 16,67 milliseconds.

2.3 Variables and Design

In order to test for the effects of credibility assessment on fixation distributions we compared two separate conditions. In both conditions subjects were asked to approach the articles as they would in real life when searching for information about an unknown topic (single-task). Half of the subjects were also told to evaluate the articles credibility while reading and after reading assign a credibility rating to the article (dual-task). The dual task condition in this study is similar to the Wikipedia screening task used by Lucassen & Schraagen (n.d.).

A pretest was conducted to identify the average time subjects needed to obtain an overview of the contents of a given article. Five subjects were given unlimited time to review three articles until they felt having a grasp of the general topic and report to the supervisor when finished.

Each subject received a different set of articles and had no influence on the choice of a topic. The mean viewing time was estimated at 4.46 minutes (SD=0.53) per article. According to these measures, time limits for the experiment were set at 5 minutes in the long time condition and 2 minutes in the restricted time condition. The 2 minutes were chosen under the assumption that they would evoke a feeling of time pressure.

Stimuli were chosen from the pool of articles classified by the Wikipedia editorial team (WET). The evaluation criteria used by the WET relate to text quality, factual accuracy, neutrality, stability, referencing[6]. In this study articles with at least "good article" status were compared to articles from the start-class. Appendix A provides a direct comparison of the criteria of the chosen article classes. Although German students were able to read the Dutch language all articles were chosen from the English Wikipedia to control for language effects. It was assumed that both Dutch and German subjects would have the most resemblance in the knowledge of the English language. Furthermore to this day, quality ratings from the WET only exist for English articles. Aside from quality criteria the choice of articles was based on an estimation of certain factors by the researcher. It was tried to balance the topics on the amount of media coverage in the past, topic difficulty, controversy, and familiarity. Subject related factors (foreknowledge, interest in the topic, personal relation with the topic) had no influence on article choices. All subjects received the same collection of articles in randomized order. Table 2.1 list the articles used in the experiment.

Table 2.1

Articles used in the experiment

Poor Quality (Start-Class)	Good Quality (at least Good Article)
Genetic Engineering	Global Warming
Superstring Theory	General Relativity
Psychics	Schizophrenia

[6] A list of the exact criteria is available on: http://en.wikipedia.org/wiki/Wikipedia:1.0/A

2.3.1 Fixation data and areas of interest

In order to compare fixation distributions article elements were divided into look zones[7]. Table 2.2 provides an overview of the article elements that were part of a certain look zone category.

Table 2.2
Look zone categories

Category	Included elements
Introduction	Title Introduction
Index	Index
Text	Text (excl. Introduction and picture subtext) Subtitles
Pictures	Pictures Info box (left of Introduction)
References	See also Notes References Further reading External links Template Categories
Other	Everything else

A list of fixation data used in our analysis was obtained via the standard export function of GazeTracker using a value of 100 ms as minimum fixation duration. Every time the gaze rested in an area of 30 by 30 pixel for longer than 100 ms a fixation was added to the output list. Using MATLAB scripts (The MathWorks, Natick, USA) fixations were allocated to look zones based on their coordinates. The comparisons of the look zone categories is complicated by the fact that categories are of differing sizes. For example the text area can be expected to receive more fixations than the index because its area is much larger. To correct for differing look zone sizes the fixation count in every look zone was divided by the total

[7] Also called areas of interest

area of all zones in a look zone category. The resulting values provide a fixation distribution as if all areas were of the same size. Finally to obtain percentage values fixation frequencies were also divided by all fixations in a trial.

To asses differences in the amount of text read on every trial the fixation frequencies inside the introduction and text areas were further analyzed. The area in both categories was subdivided into three separate equally spaced look zones: a top part, a middle part and a bottom part. Figure 2.2 shows a schematic representation of the three resulting look zones in the introduction area. To obtain percentage values the fixation frequencies were divided by the total amount of fixations in the introduction or text area respectively. Finally fixation distributions between the three parts were compared between conditions.

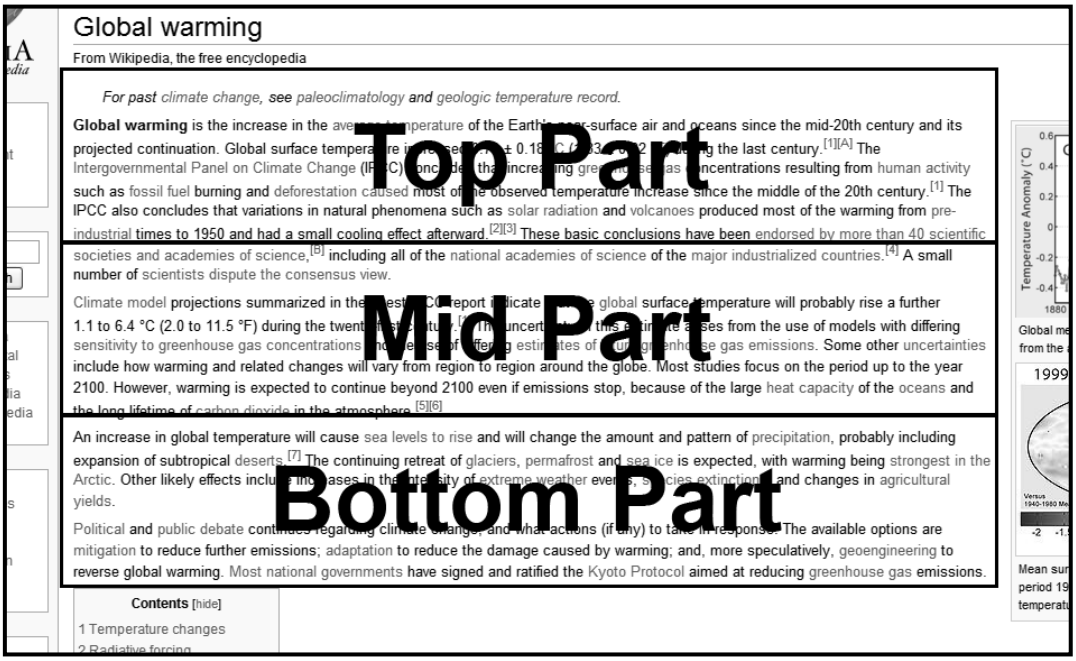

Figure 2.2 Division of the introduction into three separate look zones. The same procedure was applied on the text area.

2.3.2 Credibility ratings

Only subjects in the dual task condition were asked to rate articles for credibility and name relevant article elements. Ratings were obtained in form of a 7 point Likert-Scale and measured credibility from “not at all” (one) to “fully” (seven). In contrast, subjects in the

single-task condition were asked to rate to what extent the articles were informative , and to point out article features that they found most informative.

2.3.3 Trust cues

The frequency of trust enhancing or reducing features named by subjects in the dual-task condition was extracted from verbal reports. References to the same feature (e.g. number of pictures, quality of references, general comprehensiveness etc.) was counted only once per subject for all articles.

2.3.4 Design

The design of the experiments was a 2 × 2 × 2 repeated measures design with condition (single-task vs. dual-task) as between subjects factor and time (two minutes vs. five minutes) and article quality (good article quality vs. poor article quality) as within subjects factors. Mean fixation frequencies on article elements, mean trust ratings and categorized verbal reports of trust cues were handled as dependent variables. The order of presentation of the within subject factors was randomized among subjects.

2.4 Procedure

Subjects were welcomed and handed a sheet with information about the experiment. After reading they were asked to sign the informed consent and provide general information about themselves and their knowledge of Wikipedia in a questionnaire. Subsequently they were seated in front of the monitor under which the eye tracker was placed. The optimal distance between eyes and cameras is about 50-60 cm. Seat positions were adjusted to match this distance. After a head model creation and gaze calibration, subjects were asked if they had any remaining questions concerning the task or the tracking procedure. If not they were provided with a test article[8] that they could study for two minutes in order to get accustomed to the task, the situation and to get a feeling for the time limit. Then subjects were presented

[8] The topic of this article was „Wikipedia“ itself (version date: 23.10.2009). It was assumed that the articles content would not influence users attention distribution in following experimental articles. The article itself, although citing general criticism of Wikipedia’s open structure, does not point at particular elements that should be used to verify the information.

the six articles one at a time. Subjects were informed about the time constraint before every article . After having read an article a questionnaire was handed to the subject that asked for a informativity (respectively credibility) rating as well as for factors that had influenced this rating positively or negatively. This procedure was repeated until all articles had been viewed and rated by the subject. Finally after a short debriefing the subject was released from the experiment.

2.5 Data analysis

Due to problems with the GazeTracker software fixation data was lost in about forty percent of the recorded sessions (Appendix B gives an overview of the left data files per subject). To provide the maximum amount of usable data per subject mean frequency values of up to three articles from the two minute condition were used. The same was done for articles in the five minute condition. Fixation percentages for separate look zone categories were compared in several repeated measures ANOVA with time (short vs. long) constraint as within subjects variable and condition (single-task vs. double-task) as between subjects variable.

3. Results

3.1. Fixation distributions

3.1.1 Fixation distributions between areas of interest

Table 3.1

Percentages of fixation that fell in an area (standard deviation in parentheses)

Condition	Information		Evaluation	
Time	Short	Long	Short	Long
Introduction	53.8 (26.6)	54.5 (17.9)	63.9 (17.0)	58.1 (17.6)
Index	22.2 (22.6)	14.8 (8.6)	19.2 (15.5)	16.8 (16.1)
Text	12.2 (8.2)	18.9 (10.8)	9.4 (9.0)	13.5 (9.5)
Pictures	10.8 (16.9)	10.4 (6.8)	5.6 (7.0)	7.4 (7.3)
References	0.2 (0.6)	0.3 (0.6)	0.8 (2.0)	0.4 (0.9)

Table 3.1 gives an overview of the distribution of fixations over the predefined areas of interest. What is most obvious is a difference in fixation distribution among the areas. A 5 × 2 × 2 repeated measures ANOVA was carried out with area of interest (Introduction, Index, Text, Pictures and References) and time (short/two min. and long/five min.) as within subject variables and condition (single-task/Information and double-task/Evaluation) as between subjects variables. The dependent variable were the fixation percentages on areas of interest. Differences between areas were significant with $F(4,148)=142.0$, $p<.001$. Pairwise comparisons with Bonferroni correction showed significantly more fixations in the introduction ($M=58.3$, $SD=19.7$) than in any other area (all $p<.001$). Also fixation frequencies were significantly higher on the index ($M=18.3$, $SD=15.7$) than on pictures ($M=8.6$, $SD=9.4$) and references ($M=0.4$, $SD=1.1$) (all $p<.005$). Finally fixation percentages on references were also significantly lower than on text ($M=13.5$, $SD=9.4$) and pictures (all $p<.001$). Table 3.2 shows the mean differences of fixation percentages between areas of interest while figure 3.1 shows the overall fixation distribution. All other effects were not significant, all *F'*s<2, all *p'*s> .1. Although the within subjects effect of time was not significant ($F(1,37)=.23$, $p=.64$) as shown on figure 3.2 a trend can be seen in the effect of time on fixation distributions.

Table 3.2

Mean differences of fixation percentages between areas of interest (standard error in parentheses)

	Introduction	Index	Text	Pictures	References
Introduction	-	-	-	-	-
Index	40.0 (4.5) **	-	-	-	-
Text	44.8 (3.1) **	4.8 (2.6)	-	-	-
Pictures	49.7 (3.5) **	9.7 (2.3) *	4.9 (1.7)	-	-
References	57.9 (2.7) **	17.8 (2.0) **	13.1 (1.2)**	8.1 (1.2) **	-

Note. * p < .005, ** p < .001

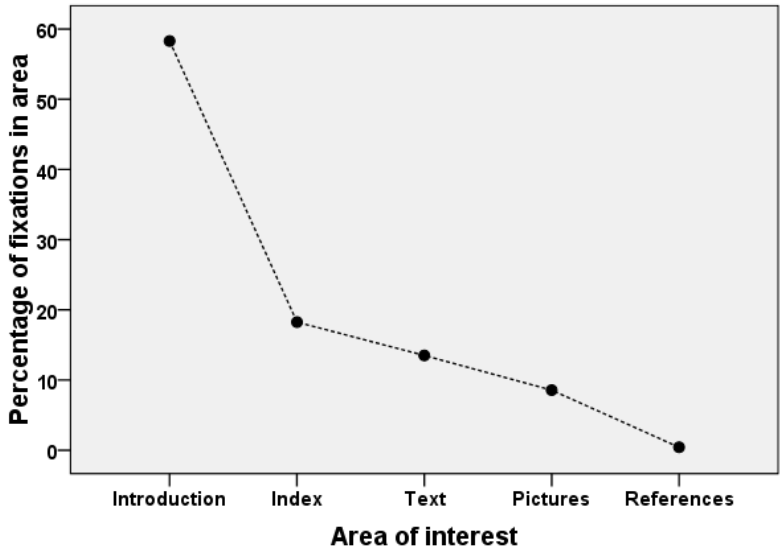

Figure 3.1 Mean fixation percentages distributed over areas of interest

These results generally confirm our first hypotheses that there is a difference in fixation distributions between areas of interest.

3.1.2. Fixation distributions in the introduction

Table 3.3

Percentages of fixation that fell in one of the three areas of the introduction (standard deviation in parentheses)

Condition	Information		Evaluation	
Time	Short	Long	Short	Long
Top part	50.0 (32.1)	44.2 (14.3)	48.5 (20.1)	48.5 (20.6)
Mid part	32.7 (24.6)	34.7 (9.9)	36.8 (17.8)	32.8 (15.4)
Bottom part	11.2 (15.3)	18.1 (10.3)	14.6 (12.8)	18.6 (8.7)

Table 3.3 gives an overview of fixation distributions inside the introduction. A 3 × 2 × 2 repeated measures ANOVA was carried out with zone location (top part, mid part and bottom part) and time (short, long) as within subject variables and condition (Information, Evaluation) as between subject variable and fixation percentages as dependent variable. The

main effect of look zone was found significant with F(2,74)=41.7, p<.001. Pairwise comparisons with Bonferroni correction showed that significantly more fixations were in the top part of the introduction (M=47.8, SD=21.5) than in the mid part (M=34.2, SD=16.6) or the bottom part (M=15.6, SD=11.3) (all p<.05). Furthermore significantly more fixations were on the mid part compared to the bottom part (p<.001). Table 3.4 shows the mean differences of fixation percentages between the areas while figure 3.2 shows overall fixation distributions. All other effects were not significant, all *F'*s<2, all *p'*s> .1.

Table 3.4

Mean differences of fixation percentages between areas of interest (standard error in parentheses)

	Top part	Mid part	Bottom part
Top part	-	-	-
Mid part	13.6 (4.3) *	-	-
Bottom part	32,2 (3.5) **	18.6 (2.5) **	-

Note. * p < .05, ** p < .001

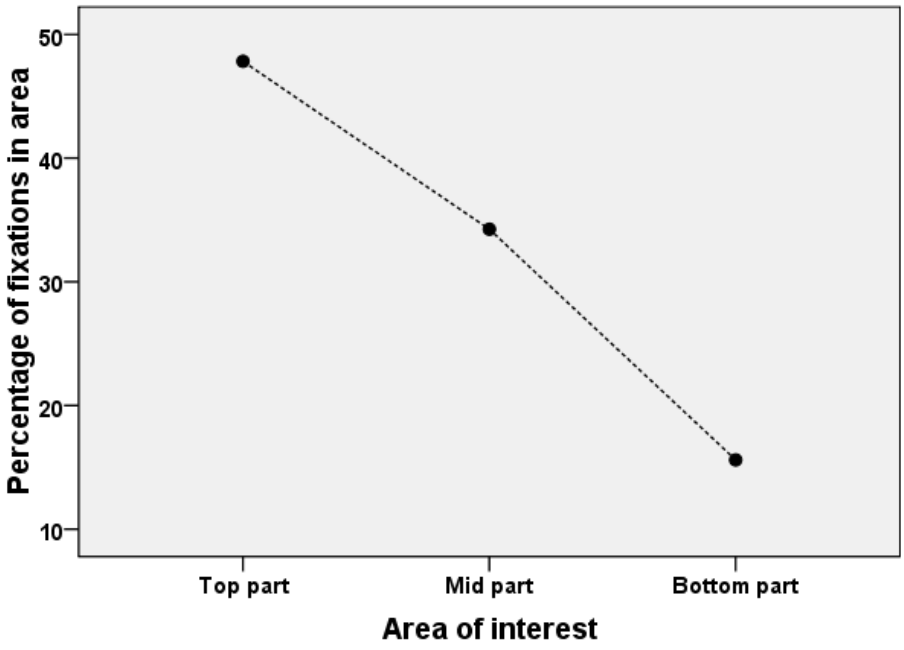

Figure 3.2 Mean fixation percentages distributed over areas of the introduction

These results show that not all text in the introduction is read. Contrary to our hypotheses this effect does not change with differing time constraints.

3.1.3. Fixation distributions in the text

Table 3.5

Percentages of fixation that fell in one of the three areas of the text (standard deviation in parentheses)

Condition	Information		Evaluation	
Time	Short	Long	Short	Long
Top part	62.6 (27.0)	51.4 (22.9)	60.3 (30.5)	61.2 (27.6)
Mid part	25.2 (22.4)	27.5 (16.8)	16.0 (16.9)	19.9 (14.7)
Bottom part	9.1 (14.9)	21.1 (17.4)	16.8 (25.3)	15.9 (19.0)

Table 3.5 gives an overview of the fixation distribution inside the text area. A 3 x 2 x 2 repeated measures ANOVA was carried out with zone location (top part, mid part and bottom part) and time (short, long) as within subject variables and condition (Information, Evaluation) as between subject variable and fixation percentages as dependent variable. The main effect of look zone was found significant $F(2,74)=69.9$, $p<.001$. Pairwise comparisons with Bonferroni correction showed that significantly more fixations in the top part of the introduction ($M=58.9$, $SD=27.2$) than in the mid part ($M=22.2$, $SD=17.7$) or the bottom part ($M=15.7$, $SD=19.8$) (all $p<.001$). No significant difference was found between the fixation frequency of the mid part compared and the bottom part ($p>.1$). Table 3.6 shows the mean differences of fixation percentages between the three areas while figure 3.3 shows the overall fixation distribution. All other effects were not significant, all *F'*s<2, all *p'*s> .1.

Table 3.6

Mean differences of fixation percentages between areas of interest (standard error in parentheses)

	Top part	Mid part	Bottom part
Top part	-	-	-
Mid part	36.7 (3.9) *	-	-
Bottom part	43.1 (4.8) *	6.4 (3.0)	-

Note. * $p < .001$

The results from this section as well as from the introductory section indicate that the upper part of a text is read more often than the rest of the text. Again this effect does not change which differing time constraints with is in stride with the hypothesis that with fewer time subjects turn away from text in favor of peripheral article information (e.g. pictures, introduction, info boxes).

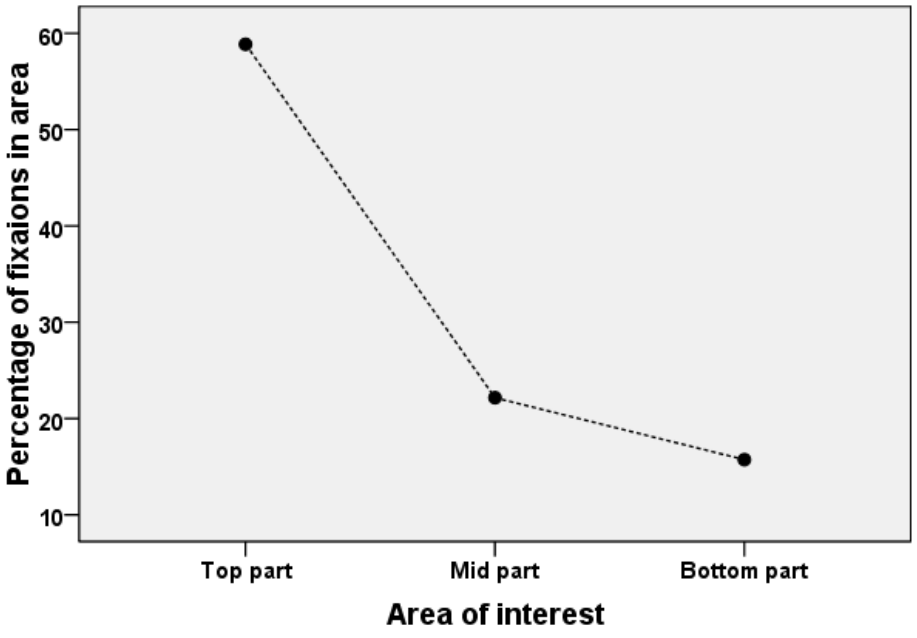

Figure 3.3 Mean fixation percentages distributed over the text area

3.2 Trust ratings

In table 3.7 trust ratings of articles in different experimental conditions are reproduced.

Table 3.7

Trust ratings for articles in different conditions (standard deviation in parentheses)

Article quality	Good		Poor	
Time constraint	Short	Long	Short	Long
Credibility rating	5.8 (1.3)	5.7 (1.0)	4.9 (1.2)	4.5 (1.7)

Note. Values are the mean of reported scores on a 7-point likert-scale (1 = *not at all credible*, 7 = *very credible*).

A significant main effect was found for article quality $F(1,18)= 9.8$, $p=.006$. Pairwise comparison with Bonferroni correction showed that good quality articles ($M=5.8$, $SD=1.1$) had a significantly higher trust rating than poor quality articles ($M=4.7$, $SD=1.4$)($p=.006$). An effect of time constraint was not significant $F(1,18)=1.0$, $p=.33$. Also a first order interaction of article quality and time constraint was not significant $F(1,18)=.25$, $p=.63$. Although the differences between Wikipedia's internal quality ratings are reflected in subjects credibility ratings there was no significant influence of time constraint on these ratings.

3.3 Verbal report frequencies

Table 3.8 gives an overview of the ten most frequent reasons named by subjects as positively or negatively influencing their credibility ratings.

Table 3.8

Frequency of factors that led to higher or lower credibility ratings.

Positive influence on rating	**frequency**	**Negative influence on rating**	**frequency**
comprehensive coverage	15	incomprehensive coverage	15
Understandable	14	not understandable	12
number of references	10	too few references	9
good structure	9	too many specialist terms	7
several viewpoints	8	bad structure	5
specialist terms	7	controversial topic	4

use of math and formulas	6	too much information	4
corresponds with own knowledge	6	biased coverage	4
seems trustworthy	6	non-scientific appearance	3
scientific appearance	5	not a scientific topic	3

Note. For the complete list see Appendix C

These frequencies show some resemblance to the percentages found by Lucassen and Schraagen (n.d.) who also found comprehensive coverage and number of references as frequently mentioned in the course of an evaluation. Note that some of the responses are not directly related to specific article elements (e.g. fit with previous knowledge, controversy of the topic).

4. Discussion

In this study we were interested in eye movement behavior of Wikipedia users while evaluating the credibility of articles. As mentioned earlier we assumed that every article element category contained information in a different form (e.g text, pictures, lists) and therefore might elicit different viewing behavior. For example factual information in text form and a schematic overview of the same information in the form of a picture leads to different fixation frequencies. Therefore we hypothesized differences in the distribution of fixation frequencies among article element categories. A comparison of those frequencies between areas of interest confirmed our hypothesis. With nearly sixty percent of overall article fixations, the article's introduction attracted the most fixations[9]. Reading is associated with short fixation durations while fixations follow each other in rapid succession with short saccades between them (Rayner, 1998). This fact may partly explain the primary position of a text element in the overall fixation frequency rating. Furthermore the introduction can be seen as a vital part for users who want to gain a general understanding of the topic. A task that both groups had to carry out while reading the articles. Querying the introduction seems to be the

[9] Remember that we controlled for element size and corrected fixation frequencies accordingly.

most straightforward strategy to achieve this goal. Looking closer we see that readers were particularly interested in the first lines of the introduction as fixation frequencies decreased from top to bottom. In general subjects seemed to be satisfied with the first few lines that often contained a brief definition of the topic. What followed in the subsequent lines was mostly general background information. Over different task and time conditions the fixation distribution on the introduction remained the same meaning that the skipping of content in the introduction happens independent of time constraint. These findings suggest that in relation to other article elements content is less relevant for information search as well as credibility assessment.

With almost twenty percent of the fixations the articles index received the second most visual attention. In the experiment the index might have helped subjects to gain a quick overview of the topic. Text elements[10] came out third with almost fourteen percent. In general text is the biggest area in all articles and would have received the most fixations if we had not corrected the data for element size. Compared to the introduction the main text goes more into the details of the topic and it contains a bigger amount of information that would be less relevant for a general understanding.

A comparison of the top, middle and bottom part again showed a primacy effect of fixations as nearly sixty percent of the fixations occurred in the top area. Subjects in the experiment were generally seen reading articles from top to bottom often reading as far as they could get within the given time. However an analysis of scan paths might be needed to further analyze text approach strategies and determine if article reading on Wikipedia is comparable to common text reading (i.e. reading a book). Given both the results of fixation distributions in the introduction and the main text it may seem strange that comprehensiveness had been most frequently named in verbal reports as an influence factor in a credibility assessment. The reason for this may lie in the different levels of comprehensiveness. The

[10] Without introduction.

findings suggests that comprehensiveness here is used on a high level describing the breadth of topics covered (i.e. deducted from an overview of the index) rather than the depth of coverage (i.e. gained from thorough reading of all available text).

Pictures received almost nine percent of overall fixations. Parts of the articles contained graphs and schematic representations of facts presented in the text so it was assumed that subjects, under time pressure would turn to faster accessible pictures instead of reading trough the text. However this was not the case as picture fixations remained constant over different time spans. An explanation of the low general fixation frequencies might be the low resolution of the images in the text which made their content rather inaccessible. For technical reasons subjects in the experiment were not allowed to click on the pictures to enlarge their content. This made the pictures especially graphs less accessible and impaired their benefit as a source of general or trust related information.

The last category were the references, which, compared to their size, received almost no fixations. Though it might be wrong to conclude that they were generally neglected. Verbal reports show that the number of references was one of the most important factors for a positive (in the case of many items) as well as a negative rating (in the case of few items). References contain very specific, source related information that is less relevant when getting a general overview of the article in an information search task. Therefore, they might not be viewed in the single task condition. In a credibility assessment they seem to be of greater importance. However, also the double-task condition did not lead to significant changes in references fixations. A reason for this might be that references are very well noticed but in a way that is not measurable using fixation frequencies. For this explanation we make a difference between reading and scanning. For example as text in an article is read or not read the fixation count rises or remains the same on a specific spot. Later this can be used as evidence that this part of the text had or had not been observed. References also contain text but users may not read it as they would do with regular text. Rather they vertically scan the

list of references for a more global estimation of its size. Metzger (2007) proposed a dual processing model of credibility evaluation were she states that when people are not motivated or not able to engage in an elaborate evaluation of the information they switch to a more heuristic evaluation of peripheral cues. Estimating the number of references can provide an idea of information quality instead of reading the references to exactly determine the origin of the information. In this case the number of references can impact a credibility assessment while their estimation is not visible in fixation frequencies. This explanation is also supported by findings from Lucassen and Schraagen (n.d.) who showed that references were among the most mentioned elements[11] of to Wikipedia users in a credibility assessment. About 25 percent of overall features mentioned in the questionnaires were related to references while almost 16 percent were about the quantity of the references list.

No evidence was found that explicitly assessing an articles credibility leads to changes in the distribution of fixations between article elements. Considering the given fixation distribution it can be assumed that subjects based a great deal of their assessment on an evaluation of elements in the text category. From verbal reports it can be inferred that subjects primarily named understandability, structure and the use of special terms as promoting or weakening a text's credibility. A more qualitative analysis of subjects scan path might reveal changes in visual approach strategies that are related to good or bad structure as well as understandability (e.g. number of recessions in eye movements, jumping between different parts in the text).

Time constraints had no significant measurable effect on fixation frequencies. It was assumed that under time constraint subjects would focus their attention away from the general text in favor of more accessible elements as pictures, index or references. However over different time constraints the distribution of fixations remained constant. Yet, this is no direct evidence that time has no effect on user behavior in a credibility assessment. As pointed out

[11] In post article questionnaires that resembled the ones used in this study.

earlier a possible effect has just not been captured with this specific method of eye fixation analysis. Again an analysis of scan paths and a more detailed analysis of text block fixations combined with retrospective think aloud protocols could provide further insights in different strategies used by subjects under different conditions.

Finally credibility ratings reflected article quality ratings assessed by the Wikipedia editorial team which confirms our hypotheses. Although mean differences between ratings have been less distinctive as we assumed when comparing articles of fairly high and low categories within Wikipedia's internal rating system. Even though articles in the low quality condition oftentimes missed adequate referencing, useful pictures and comprehensive coverage they were given an average rating of 4.7 on a seven point likert-scale. Also the ratings showed no effect of time constraint, an argument against the assumption that time acts as an amplifier for credibility. Time may well have an influence on the strength and the belief in the correctness of a certain rating. According to dual-processing models (e.g. Petty & Caciappo, 1981) time may enable subjects to engage in a more elaborate processing of the information by that strengthening the faith in their final judgment. Another question that results from the missing effect of time constraint is whether credibility ratings are formed at the time of the assessment through iterative processes as proposed by Hilligoss and Rieh (2007) or afterwards as an holistic overall estimation and to what extent these ratings are based on actual article features or the feeling of credibility that is previously assumed by a subject given a certain topic. For example several subjects remarked not trusting the information in the article about psychic's. When asked about the reasons it became clear that they disregarded and were suspicious of paranormal phenomenon's in general. Failing to realize that even information presented in an article on a controversial topic can be perfectly credible (and needs to be judged independently from the topic) given that the article is well written.

Because of experimental reasons clicking on the links in an article was prohibited. Some subjects reported that this would deviate from their normal reading behavior on Wikipedia as they would click on words they did not understand, and read the general definition, before they continued reading the initial article. Because of technical reasons scrolling with the mouse wheel was also not allowed. Subjects had to make use of Internet Explorers scroll bar which was found unusual and disturbing by some as they had to focus away from the text to click on the scroll bar and then refocus on the article. As stated before it was hard for subjects to include pictures in to their evaluations of the article. In the used articles some pictures are shown in a smaller size to fit the article and are therefore only visible in lower resolution. Therefore some pictures, especially graphs or schematic overviews were not accessible. As a compensation the text under the pictures was counted as belonging to the picture as it enabled the subject to make assumptions and gain clarity about the content of the pictures. So interest in picture subtext counts as interest in the picture and not in the text.

This study provided an exploratory overview on fixation distributions in a credibility assessment. Further research should go into detailed analysis of how specific article elements affect credibility and are treated by subjects. Also more attention needs to be spent on choosing good articles as experimental stimuli (e.g. controlled for topic, controversy, quality, correctness, popularity, importance, actuality etc.). Other important topics are the influence of other context factors (e.g. high vs. low risk situation) and subject factors (e.g. foreknowledge of the topic, experience with Wikipedia, internet use, experience with social science methodology etc.). Finally the relation between credibility and trust needs to be further analyzed for the context of Wikipedia. A subject in a study by Richman and Wu (2008) hits the spot by stating:" I have never doubted the credibility of a particular page, but I am somewhat dubious about completely trusting the site as a whole".

In conclusion the present study is the first to look at visual attention of Wikipedia users in a credibility assessment. Fixation distributions clearly vary between article categories and show a primacy effect in text elements. No interaction with time constraint and task was found. Interestingly the analysis of fixation distribution frequencies lead to different conclusions about the importance of article elements than verbal reports found in this study and earlier research by Lucassen and Schraagen (n.d.). More research is needed to study this relationship between verbal reports an eye movement data in the context of a credibility assessment.

Aknowledgements

I would like to thank all those who helped me in the proceeding and the realization of this thesis. Special thanks go to Prof. Dr. Jan Marten Schraagen for the inspiration to work on trust in Wikipedia, Teun Lucassen for his help and insights on everything related to Wikipedia, and finally Dr. Matthijs Noordzij for his supervision and encouragement well beyond this thesis research alone.

REFERENCES

Alexa Internet, Inc. (2010, January 30). *Top sites* [Data file]. Retrieved from http://s3.amazonaws.com/alexa-static/top-1m.csv.zip

Beymer, D., Orton, P.Z., Russell, D.M. (2007). An eye tracking study of how pictures influence online reading. *Interact, 2*, 456-460.

Chen, L. (2007, March 28). Several colleges push to ban Wikipedia as resource. *The Duke Chronicle.* Retrieved from http://dukechronicle.com/node/142611

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58,* 737–758.

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM, 48*(12), 152-152.

Duchowski, A.T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments and Computing, 34*(4), 455-470.

Fitts, P.M., Jones, R.E. & Milton, J.L. (1950). Eye movement of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review, 9,* 24-29.

Flanagin, A.J., & Metzger, M.J. (2000). Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly, 77*(3), 515–540.

Fogg, B.J. and Tseng, H. (1999). The elements of computer credibility. *CHI '99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 80-87), New York: ACM Press.

Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., & Treinen, M. (2001). What makes a web site credible? A report on a large quantitative study. *Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems* (pp. 66-68), Seattle: ACM Press.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature, 438,* 900-901. Retrieved from http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

Halavais, A. (2004, August 29). The Isuzu Experiment [Blog]. Retrieved from http://alex.halavais.net/the-isuzu-experiment

Hilligoss, B., & Rieh S.Y. (2007). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management, 44*(4), 1467-1484.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Lucassen, T. & Schraagen, J.M. (n.d.). *Trust in Wikipedia: How users trust information from an unknown source.* Submitted to WICOW 2010. Department of Cognitive Psychology and Ergonomics, University of Twente. Retrieved from Teun Lucassen.

Lim, S. (2009). How and why do college students use Wikipedia. *Journal of the American Society for Information Science and Technology, 60*(11), 2189-2202.

Magnus, P.D. (2008). Early response to false claims in Wikipedia. *First Monday, 13*(9).

McGuinness, D.L., Zeng, H., da Silva, P.P., Ding, L., Narayanan, D., & Bhaowal, M. (2006). Investigations into trust for collaborative information repositories: A Wikipedia case study. *WWW 2006 15th International World Wide Web Conference*. Edinburgh, UK.

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58*(13), 2078–2091.

Mills, D. L. (1991).Internet time synchronization: The Network Time Protocol. IEEE *Transactions on Communications, 39(* 10), 1482-1493.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004). The determinants of web page viewing behavior: an eye-tracking study. *ETRA '04: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications,* (pp. 147-154), New York, NY, USA. ACM Press.

Petty, R. and Cacioppo, J.T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology, 19,* 124-205.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 371-422.

Richman, J. and Wu, L. (2008). Visual representations on Wikipedia: Less is more. Unpublished manuscript. Stanford Visualization Group, Stanford University.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651-665.

Scholz-Crane, A. (1998). Evaluating the future: A preliminary study of the process of how undergraduate students evaluate Web sources. *Reference Services Review, 26*(3/4), 53-60.

Simons, D.J. (2000). Attentional capture and inattentional blindness. Trends in Cognitive Sciences, 4(4), 147-155.

Simons, D.J. & Levin, D.T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261-267.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), *Eye Movements and their Role in Visual and Cognitive Processes* (pp. 353-393). Amsterdam: Elsevier.

Wang, Y. D. & Emurian, H. H. (2005). An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior, 21*, 105-125.

Wathen, N. C. & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology, 53*(2), 134-144.

Wikipedia:Size_comparisons (2010, January 29). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

APPENDIX A - QUALITY CRITERIA FOR ARTICLES COMPARED IN THE EXPERIMENT

Start-Class	Good-Article-Class
1. Developing 2. Has a usable amount of good content but is weak in many areas. 3. Still incomplete 4. Lacks adequate reliable sources an referencing 5. Provides enough sources to establish verifiability 6. Quality of the prose may be distinctly unencyclopedic 7. Should not be in any danger of being speedily deleted.	1. Well-written: - Prose is clear and the spelling and grammar are correct - complies with the manual of style guidelines for lead sections, layout, jargon, words to avoid, fiction, and list incorporation. 2. Factually accurate and verifiable: - provides references to all sources of information in the sections dedicated to the attribution of these sources according to the guide to layout - provides in-line citations from reliable sources for direct quotations, statistics, published opinion, counter-intuitive or controversial statements that are challenged or likely to be challenged, and contentious material relating to living persons - contains no original research 3. Broad in its coverage: - addresses the main aspects of the topic - stays focused on the topic without going into unnecessary detail 4. Neutral: - represents viewpoints fairly and without bias 5. Stable: - does not change significantly from day-to-day because of an ongoing edit war or content dispute. 6. Illustrated, if possible, by images: - images are tagged with their copyright status, and valid fair use rationales are provided for non-free content - images are relevant to the topic, and have suitable captions.

From: http://en.wikipedia.org/wiki/Wikipedia:1.0/A

APPENDIX B - RECORDED SESSIONS PER SUBJECT PER TRIAL

Subject	1	2	3	4	5	6	total
2	1	1	1	1	1	1	**6**
3	1	1	1	1	1	1	**6**
4	1	1	0	0	0	1	**3**
5	1	0	0	0	1	0	**2**
6	0	0	1	1	1	1	**4**
7	0	1	1	1	0	0	**3**
8	0	1	0	0	1	1	**3**
9	1	1	0	1	1	0	**4**
10	1	1	0	1	1	0	**4**
11	1	0	1	1	1	1	**5**
12	1	1	1	1	1	0	**5**
13	1	1	1	0	0	1	**4**
14	1	1	0	0	1	0	**3**
15	1	0	1	1	1	1	**5**
16	0	1	1	1	0	0	**3**
17	1	0	0	1	1	0	**3**
18	1	1	0	1	0	0	**3**
19	1	1	0	0	0	1	**3**
20	1	1	1	1	0	1	**5**
21	1	1	1	0	0	1	**4**
22	0	0	1	1	1	1	**4**
23	1	0	1	1	0	0	**3**
24	1	1	1	1	0	1	**5**
25	1	1	0	1	0	0	**3**
26	0	0	1	1	1	0	**3**
27	1	1	1	0	0	1	**4**
28	0	1	1	0	1	0	**3**
29	1	0	0	1	1	0	**3**
30	0	1	0	1	1	0	**3**
31	0	1	1	0	0	0	**2**
32	1	1	0	1	1	1	**5**
33	1	0	0	1	0	1	**3**
34	1	0	0	1	1	1	**4**
35	1	1	1	0	1	0	**4**
36	0	0	1	1	0	1	**3**
37	0	1	0	0	0	1	**2**
38	0	0	0	1	0	1	**2**
39	1	0	0	1	1	1	**4**
40	0	0	1	0	0	0	**1**
41	0	0	0	0	1	1	**2**
42	0	1	0	0	0	1	**2**
43	1	1	0	1	0	1	**4**
44	0	1	1	0	1	1	**4**
Total	**27**	**27**	**22**	**27**	**23**	**25**	**151**

APPENDIX C – FACTORS THAT INFLUENCED CREDIBILITY RATINGS

Positive influence on rating	**frequency**
comprehensive coverage	15
understandable	14
number of references	10
good structure	9
several viewpoints	8
specialist terms	7
use of math and formulas	6
corresponds with own knowledge	6
seems trustworthy	6
scientific appearance	5
use of pictures	4
refers to popular scientists	4
examples given	4
calls on facts	4
contains many numbers	3
use of diagrams	2
article length	2
good internal consistency	1
not too long	1
no controversial topic	1
Many internal links	1
many notes	1
actuality	1

Negative influence on rating	**frequency**
incomprehensive coverage	15
not understandable	12
too few references	9
too many specialist terms	7
bad structure	5
controversial topic	4
too much information	4
biased coverage	4
non-scientific appearance	3
not a scientific topic	3
truth cannot be proven	3
unknown topic	2
useless pictures	2
too manny notes	1
outdated information	1
contradicts own opinion	1
not interesting	1