Evaluating WikiTrust: A trust support tool for Wikipedia

by Teun Lucassen and
Jan Maarten Schraagen

Abstract

Because of the open character of Wikipedia readers should always be aware of the possibility of false information. WikiTrust aims at helping readers to judge the trustworthiness of articles by coloring the background of less trustworthy words in a shade of orange. In this study we look into the effects of such coloring on reading behavior and trust evaluation by means of an eye-tracking experiment. The results show that readers had more difficulties reading the articles with coloring than without coloring. Trust in heavily colored articles was lower. The main concern is that the participants in our experiment rated usefulness of WikiTrust low.

Contents

- [1. Introduction](#)
 - [2. Methodology](#)
 - [3. Results](#)
 - [4. Discussion and conclusions](#)
-

1. Introduction

The high quality of encyclopedic content on Wikipedia is becoming more and more widely recognized. A recent study demonstrated that Wikipedia's quality is comparable to *Encyclopædia Britannica* (Giles, 2005).

However, one of the greatest advantages of Wikipedia is perhaps one of its biggest weaknesses, namely its open character. Everyone with an Internet connection can edit the articles, mostly even without having to login. This is the main reason for its enormous growth since its launch in 2001, but this also means that readers always have to be aware of false information. Magnus (2008) showed that the Wikipedia community is quick to correct deliberate errors, but not all errors were quickly noticed and corrected.

Usually, when studying information trust or credibility, the source of information is very important (Leggatt and McGuinness, 2006). Imagine that you are in the market for a new car. When you come across a new piece of information on a specific model, you want to know the source of this information. You will interpret it differently if it is an advertisement by the manufacturer rather than a review by an independent car magazine.

On Wikipedia, the strategy to assess the source of information to estimate trustworthiness cannot be applied, because this information is not available. You mostly only have an IP address or username about a given author. With those articles with a high number of contributors, it may be impossible to assign authority or responsibility to a single person or institution. Normal heuristics do not apply and other ways are needed to assess trustworthiness. A think-aloud experiment has revealed some of the strategies that some Wikipedia readers use (Lucassen and Schraagen, 2010). In this experiment, it was shown that besides trying to verify the factual accuracy of information, readers also looked at several heuristic cues, such as the number of references or images.

The fact that other, new ways of assessing trustworthiness need to be taken infers that this task is difficult for an average user. Several attempts to support readers in assessing trustworthiness have been taken. McGuinness, *et al.* (2006) designed an algorithm that calculated a trust value based on internal links to an article. Internal links are links to other Wikipedia articles, recognizable as blue, underlined words in the text. The algorithm is based on the assumption that when an internal link in another article is made to a specific article, it is a token of trust. When the topic of this article is mentioned, but no internal link is created,

this is a token of distrust. Trust can be calculated by the link ratio on an author level and article level.

Kittur, *et al.* (2008) proposed a system which visualized the edit history of Wikipedia articles. This system provided insight into how each author contributed to a given article. For example, one author might be responsible for most of the text in one article, whereas another article consists of contributions from a vast number of authors. How to interpret these observations is left to the user.

Another approach was taken by Cross (2006). He proposed a system in which passages that are presumably less trustworthy were colored. In this system, trustworthiness was measured by the age of the text. New text was colored red (untrustworthy), slighter older text was yellow, even older text was green, and the oldest text was black (trustworthy). This system relied on an assumption that errors in an article will sooner or later be noticed and corrected by someone, making older text more trustworthy. However, the validity of the use of edit age has been questioned by Luyt, *et al.* (2008). They argued that new edits were mainly focused on adding new information (which may contain new errors) instead of correcting existing information.

Nevertheless, a promising approach was taken by Adler, *et al.* (2008) in their WikiTrust system. Trust was determined by the survival duration of each single word. The background of each word was colored using a shade of orange ranging from white for trustworthy and dark orange for untrustworthy. An example of such background coloring is shown in [Figure 1](#). Consider when an article is edited by someone. When this person leaves a certain portion of the text unchanged, an indirect vote is given to the words in this text. This increases the trustworthiness value of these words. Newly added words receive the trustworthiness of its author which is determined by the average survival duration of his added words.

Figure 1: A Wikipedia article colored by WikiTrust.

Note that this explanation is simplified for clarity. Several other factors, such as how to cope with reverting edits, are left out here. Also, edits further away from a certain text (in words) will increase the trustworthiness of this text less than closer edits. For a formal description of the system, see Adler, *et al.* (2008).

An important feature of WikiTrust is that it does not estimate the trustworthiness of an article as a whole. Instead it points out sections of articles which have recently been changed and are expected to be less trustworthy. This way, a recent, untrustworthy edit does not necessarily harm the trustworthiness of an entire article. Problems in one particular section of an article can easily be spotted. WikiTrust has become available on live Wikipedia in several languages through a Mozilla Firefox plug-in [[1](#)] in early 2010. This plug-in adds a trust tab to Wikipedia which shows the colored version of the page.

Little is known about the effects of WikiTrust on lay readers. First, it can be argued that reading a colored article is more difficult than the original. This is not directly related to trust in the article, but it can be distracting. Important questions that need to be answered are: Is perceived trustworthiness influenced by the system? Is the system useful?

This study focuses on the behavior of readers while using WikiTrust. Naïve participants were presented several Wikipedia articles with and without WikiTrust text coloring. They were asked to read the articles to understand their content and assess its trustworthiness. Behavior was measured using eye-tracking and post-trial questionnaires.

Eye-tracking measures are a good indicator of reading difficulties (Rayner, 1998). Re-reading leads to a higher number of fixations (nearly motionless gaze) in a text. Rinck, *et al.* (2003) found that the appearance of conflicting information within a text led to re-reading and thus greater fixations.

Reading difficulties can also lead to longer durations of fixations. Next to these specific eye-tracking observations, longer total reading times can be expected with reading difficulties.

Background coloring may be influential on reading behavior on two levels, namely perceptual and cognitive. First, the reduced contrast of colored words in comparison to words with a white background will make them more difficult to read. Further, we expect colored words to draw bottom-up and top-down attention. Colored words will stand out from the rest of the text and will draw attention. Readers will also be aware of the purpose of the coloring system and therefore draw attention to these words, trying to incorporate their added value. Voluntary and

involuntary shifts of attention to colored words are likely to interrupt natural reading. This leads to the following hypothesis:

Hypothesis 1: Reading a text with WikiTrust coloring is more difficult than without coloring.

It has to be noted that an important feature of WikiTrust is that it comes in a separate tab on a given Wikipedia article. This means that it is also possible to read the article without coloring first before looking at trustworthiness.

The desired effect of WikiTrust is that readers are aware of possible problems concerning the trustworthiness of Wikipedia articles. We expect that when WikiTrust points out major recent changes by showing a heavily colored text, trust in this article will be lower than when WikiTrust is not enabled. This leads to the following hypothesis:

Hypothesis 2: Readers have less trust in articles which are heavily colored by WikiTrust compared with the same articles without WikiTrust.

Further, we expect that because of the intuitive but sophisticated algorithm used in WikiTrust, readers are able to see the benefits from using the support system. This leads to the following hypothesis:

Hypothesis 3: Readers recognize the added value of WikiTrust.

2. Methodology

2.1. Participants

Fourteen college students (nine male, five female) aged between 18 and 27 ($M=23.14$, $SD=2.14$) took part in this experiment as volunteers. All participants reported normal or corrected to normal vision. None of the participants reported color-blindness. The participants were of Dutch or German nationalities and were proficient in English used in the stimuli. College students are particularly suitable for research on behavior on Wikipedia, because they are an important group of regular users (Lim, 2009). All participants had experience with Wikipedia and could explain the basics of the Web site in their own words.

2.2. Task

The task performed in this experiment was twofold. Participants had to read the introduction of a Wikipedia article for general information comprehension. Next, participants were asked to assess the trustworthiness of the presented information. This task is similar to the Wikipedia Screening Task as introduced by Lucassen and Schraagen (2010).

2.3. Independent variables

2.3.1. Intensity of coloring

Three different articles containing three levels of WikiTrust coloring were presented to each participant. One article was not colored, one article contained light coloring (10–15 percent of the text), and one article was heavily colored (90–98 percent of the text). The light coloring condition reflects the situation in which some small parts of the article have been edited recently, but in which WikiTrust marks the article as quite reliable. The heavily colored condition reflects the situation in which the article has undergone major recent changes and in which WikiTrust marks the article as unreliable.

The articles in this experiment were on the topics 'Banana', 'Big Bang', and 'Diabetes mellitus'. Their quality was reasonably high, respectively receiving the rating C-class, Featured article and B-class by the Wikipedia Editorial Team [2]. The three color intensities were created manually for each article using the 'multiplication' level filter in Adobe Photoshop 8. Each participant received three different articles with three different coloring intensities. Every combination of topic and color intensity was presented an equal number of times.

2.4. Dependent variables

2.4.1. Reading behavior

Reading behavior of the participants was measured by the number of fixations, the average fixation duration and the total reading time. The number of fixations was corrected for the length of the introduction to reflect an average introduction length of 500 words. The same was done for reading time.

2.4.2. Trust assessment

The effect of the WikiTrust system on the perceived trustworthiness of the articles was measured using a questionnaire after each article. This questionnaire contained seven-point Likert scales (with only the end points labeled) on perceived trustworthiness of the article (Trust), perceived interference of WikiTrust with their own assessment (Interference), usefulness of WikiTrust (Usefulness), and perceived readability of the article (Readability).

2.5. Procedure

Participants were positioned in front of a 17" CRT screen, fitted with a FaceLab 4.5 eye-tracker. This system consists of two FireWire cameras and an infrared LED light located under the screen. After calibration, participants were given written instructions on the task in the experiment, Wikipedia, and WikiTrust. Instructions on WikiTrust were similar to the original instructions as provided by the actual system.

Each participant performed the task in three trials with unlimited time. After each trial a questionnaire on the trust evaluation was presented. The experiment ended after about 30 minutes with a debriefing on the purpose of the study.

3. Results

3.1. Reading behavior

Table 1 shows the results for the number of fixations, fixation durations and reading times for the three coloring conditions.

Table 1: Reading behavior results. Note: M=mean, SD=standard deviation.						
	No coloring		Light coloring		Heavy coloring	
	M	SD	M	SD	M	SD
Number of fixations	523	109	680	219	628	150
Fixation duration (ms)	232	44	229	39	269	63
Reading time (sec)	125	46	161	56	167	56

A significant difference was found between the number of fixations in the no coloring and light coloring condition; $t(11)=2.25$, $p=.046$. For the difference in the number of fixations in the no coloring and heavy coloring condition a trend was found; $t(11)=2.00$, $p=.071$. No difference was found between both colored conditions.

The difference in fixation duration was significant for the no coloring and heavy coloring condition; $t(11)=4.40$, $p=.001$, and for the light coloring and heavy coloring condition; $t(11)=3.36$, $p=.006$. No difference was found between the no coloring and light coloring condition.

Reading times were significantly longer in the heavy coloring condition than in the no coloring condition; $t(11)=3.85$, $p=.003$. A trend was found between the no coloring and light coloring condition; $t(11)=2.04$, $p=.066$. No difference was found between both colored conditions.

Based on the significant differences between the no coloring condition and both colored conditions the first hypothesis is accepted.

3.2. Trust assessment

Table 2 shows the results of the four items on trust assessment in the posttrial questionnaires.

Table 2: Trust assessment results. Note: 1=lowest, 7=highest, M=mean, m=median.						
	No coloring		Light coloring		Heavy coloring	
	M	m	M	m	M	m
Trust	4.92	5	5.42	6	4.00	4
Interference	n/a	n/a	3.67	3	4.93	5
Usefulness	n/a	n/a	3.50	3	2.43	2

Readability	n/a	n/a	4.00	4.5	1.79	1.5
-------------	-----	-----	------	-----	------	-----

As the Likert scales are assumed to be measuring at an ordinal rather than an interval level (Jamieson, 2008), Wilcoxon Signed Rank Tests were performed, rather than parametric t-tests.

Trust was significantly lower in the heavy coloring condition than in both other conditions (no coloring: $Z=2.13$, $p=.033$; light coloring: $Z=2.06$, $p=.040$). No difference was found between the no coloring and light coloring condition.

Based on the observation that trust is influenced by WikiTrust coloring, we accept the second hypothesis.

Interference was significantly higher in the heavy coloring condition than in the light coloring condition ($Z=2.48$, $p=.013$). Comparison with the no coloring condition is not useful, since no interference was possible in this condition.

Usefulness was rated higher in the light coloring condition than in the heavy coloring condition ($Z=2.22$, $p=.026$). Usefulness was not rated in the no coloring condition.

Readability was better in the light coloring condition than in the heavy coloring condition ($Z=2.95$, $p=.003$). Readability in the no coloring condition was not taken into account since the text was not influenced by WikiTrust. These observations on readability contribute to accepting Hypothesis 1.

Based on the low scores on usefulness, Hypothesis 3 has to be rejected. However, it is dependent on the intensity of the coloring.

4. Discussion and conclusions

In this study we explored the effects of WikiTrust on the reading behavior and trust assessment of readers. The hypothesis that WikiTrust hampers the reading performance of readers was confirmed. This was not only observable in the eye-tracking data, but also noticed by the participants (Readability). Trust ratings were influenced by the coloring of WikiTrust. However, the added value of the system is limited. Usefulness was rated about three out of seven points.

It is important to note that the reading difficulties evoked by WikiTrust are only a limited predicament. The system has to be enabled by selecting the 'Trust tab', so normal reading for information can also be done without text coloring. Of course, when doing this, no trust information is provided to the user. This information can be accessed afterwards, but it is questionable whether users are motivated to do so. A dual processing model of credibility evaluation has been proposed which predicts when people will engage in an elaborate systematic evaluation (Metzger, 2007). This is based on both the ability and motivation of the user. Ability will be higher using WikiTrust, but motivation will likely be low. It has been shown that people are aware of important cues they could use in such trustworthiness evaluations, but hardly use them (Walraven, *et al.*, 2009).

A different issue is raised when we assume that readers are in fact using the system. This study has showed that the perceived trustworthiness is influenced by the presence of the WikiTrust system: Articles were considered less trustworthy when heavily colored compared with the same articles without coloring. This means that WikiTrust was not simply ignored by the participants and the coloring had the desired effect on the judgments of the participants.

However, our main concern is that usefulness was rated low by the participants. WikiTrust seems to be slightly more useful when only small parts are colored, but even then usefulness is limited. Participants in our experiment were not sure what to do with the information on the age of words in the text. Further development of WikiTrust could benefit from knowledge about the (heuristic) strategies of Wikipedia users when assessing trustworthiness.

WikiTrust is a promising support tool and in fact the only one that made it to the stage where it is actually available to the Wikipedia public. This study has shown that the decision to present trust information by a separate tab was right since reading behavior is affected by its coloring. However, more effort should be put into the usability of the system. This study showed that users are having problems to see how they can benefit from it, even though a clear explanation was provided and the participants were highly educated Master's students.

About the authors

Teun Lucassen is Ph.D. candidate in the Department of Cognitive Psychology & Ergonomics at the University of Twente, the Netherlands. He received his Master's degree in the Department of Human Media Interaction at the same university. Currently, Teun is researching trust in collaborative repositories (such as Wikipedia) for his Ph.D. dissertation. Direct comments to `t [dot] lucassen [at] gw [dot] utwente [dot] nl`

Jan Maarten Schraagen, is Senior Research Scientist at TNO Human Factors and Professor of Applied Cognitive Psychology at the University of Twente. His research interests include task analysis, team decision-making, trust in collaborative repositories, and adaptive human-computer collaboration. He was main editor of *Cognitive task analysis* (edited with Susan F. Chipman and Valerie L. Shalin; Mahwah, N.J.: Lawrence Erlbaum Associates, 2000) and *Naturalistic decision making and macrocognition* (Aldershot, England: Ashgate, 2008). Dr. Schraagen holds a Ph.D. in cognitive psychology from the University of Amsterdam, the Netherlands.

Acknowledgments

We would like to thank Dominic Portain, Marc Lauffs, Malte Risto and Jacek Sliwinski for their efforts in gathering data by means of an experiment.

Notes

1. <https://addons.mozilla.org/en-US/firefox/addon/11087/>, accessed 15 April 2011.

2. http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment, accessed 15 April 2011.

References

Bo Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman, 2008. "Assigning trust to Wikipedia content," *WikiSym '08: Proceedings of the 4th International Symposium on Wikis* (8–10 September, Porto, Portugal).

Tom Cross, 2006. "Puppy smoothies: Improving the reliability of open, collaborative wikis," *First Monday*, volume 11, number 9, at <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1400/1318>, accessed 12 July 2010.

Jim Giles, 2005. "Internet encyclopaedias go head to head," *Nature*, volume 438, number 7070, pp. 900–901.

Susan Jamieson, 2004. "Likert scales: How to (ab)use them," *Medical Education*, volume 38, number 12, pp. 1,217–1,218.

Andrew Leggatt and Barry McGuinness, 2006. "Factors influencing information trust and distrust in a sensemaking task," *Proceedings of the 11th International Command and Control Research and Technology Symposium* (Cambridge, U.K.; 26–28 September), at http://www.dodccrp.org/events/11th_ICCRTS/html/papers/156.pdf, accessed 15 April 2011.

Brendan Luyt, Tay Chee Hsien Aaron, Lim Hai Thian, and Cheng Kian Hong, 2008. "Improving Wikipedia's accuracy: Is edit age a solution?" *Journal of the American Society for Information Science and Technology*, volume 59, number 2, pp. 318–330.

Teun Lucassen and Jan Maarten Schraagen, 2010. "Trust in Wikipedia: How users trust information from an unknown source," *WICOW '10: Proceedings of the 4th Workshop on Information Credibility*, pp. 19–26.

P.D. Magnus, 2008. "Early response to false claims in Wikipedia," *First Monday*, volume 13, number 9, at <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2115/2027>, accessed 12 July 2010.

Deborah L. McGuinness, Honglei Zeng, Paulo Pinheiro da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal, 2006. "Investigations into trust for collaborative information repositories: A Wikipedia case study," *WWW2006: Proceedings of the 15th International World Wide Web Conference* (Edinburgh; 22–26 May).

Miriam J. Metzger, 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*, volume 58, number 13, pp. 2,078–2,091.

Keith Rayner, 1998. "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, volume 124, number 3, pp. 372–422.

Mike Rinck, Elena Gámez, José M. Díaz, and Manuel De Vega, 2003. "Processing of temporal information: Evidence from eye movements," *Memory & Cognition*, volume 31, number 1, pp. 77–86.

Amber Walraven, Saskia Brand–Gruwel, and Henny P.A. Boshuizen, 2009. "How students evaluate information and sources when searching the World Wide Web for information," *Computers & Education*, volume 52, number 1, pp. 234–246.

Editorial history

Received 12 July 2010; accepted 12 April 2011.

This paper is licensed under a [Creative Commons Public Domain License](https://creativecommons.org/publicdomain/zero/1.0/).

Evaluating WikiTrust: A trust support tool for Wikipedia
by Teun Lucassen and Jan Maarten Schraagen.

First Monday, Volume 16, Number 5 - 2 May 2011

<http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/prINTERfriendly/3070/2952>