ARTICLE FEEDBACK TOOL: SUPPORTING CREDIBILITY EVALUATIONS ON WIKIPEDIA

BACHELOR THESIS BY WELMOED LOOGE (S0208841)

Supervisors: Dr. Teun Lucassen & Prof. Dr. Jan Maarten Schraagen

21-08-2013

ABSTRACT

The Internet allows users to upload or edit information anonymously. One of the most well-known examples of this principle is the online encyclopedia Wikipedia. Because the true source of the information is hard to determine, users often do not know how to evaluate credibility of the information. A possible solution to help users evaluate the credibility of information is through the use of a decision support system, which can provide the user with an advice about the credibility of the information. This thesis focuses on the Article Feedback Tool, a rating system that is being developed by Wikipedia to improve article quality. This tool could potentially function as a decision support system, and therefore this thesis studies whether it influences people's credibility evaluations. Furthermore, we investigated whether the placement of the Article Feedback Tool has any effect on credibility evaluations. In the experiment, participants were asked to read Wikipedia articles and to answer questions about them and about their trust in the articles. Analysis showed no influence of the Article Feedback Tool on the participants' credibility evaluations at all. Therefore, we argue that the Article Feedback Tool is not used as support system in its current form nor in the more visually prominent form that we tested in the experiment. However, all participants were academic students and they generally were able to distinguish between high-quality and low-quality information based on their information skills. This may indicate that people who possess information skills have no need for a support system in evaluating credibility. We therefore recommend that future research should study whether or not the Article Feedback Tool does support credibility evaluations in people with less trained information skills.

TABLE OF CONTENTS

INTRODUCTION

Due to the open structure of the Internet, anyone is able to put information online. A website that is known for its open-editing structure is the free encyclopedia Wikipedia. The fact that anyone can contribute and edit articles anonymously calls for a careful evaluation of information quality and credibility. However, research shows that people often do not know how to evaluate credibility (Lucassen & Schraagen, 2011b), and that the people that do know often do not do so (Walraven, Brand-Gruwel & Boshuizen, 2009). To aid people in this decision making process, a support system could be used. Wikipedia is developing the 'Article Feedback Tool'(AFT), a tool that aims to improve article quality and to get more users involved in editing and writing articles. This tool could also be used as a support system to help people in their credibility evaluations by providing them with an advice on the article quality. The current version of the AFT (version 4) is located at the bottom of every article, right below the references, but literature suggests that prominence is a necessary factor for credibility evaluations (Fogg, 2003). In this study, we will investigate whether or not the AFT, and its position on the web page, is of influence on people's credibility evaluations.

We will begin by defining credibility and describing the way people generally evaluate credibility of (online) information and what role prominence plays in this process. After that, several ways to support people's credibility evaluations will be described, followed by the hypotheses and the experiment to put those to the test. Finally, the results will be presented and discussed.

EVALUATING CREDIBILITY

Before explaining the process of evaluating credibility, we will define credibility and a closely related concept: trust. According to Hovland and Weiss (1951), Hovland, Janis and Kelley (1953), Fogg and Tseng (1999) and Self (2009) credibility boils down to believability, and consists of two essential components: trustworthiness and expertise. They describe trustworthiness as "the perceived goodness or morality of the source", while expertise is explained as "the perceived knowledge and skill of the source". Information is thus deemed credible if it is believable, and the source is both knowledgeable and trustworthy.

A definition of trust is given in McKnight and Chervany (1996), where trust is described as a person's willingness to depend on somebody or something even though they

risk negative consequences by doing so. According to Lucassen and Schmettow (2011), trusting information involves a risk of trusting low-quality information. This risk can be diminished by evaluating the credibility of information.

The need for credibility evaluations stems from the open structure of the Internet, which allows anyone to put information online. A textbook example of the open structure of the Internet is the online encyclopedia Wikipedia. The open-editing structure that underlies the website has the advantage that articles can be easily generated and are updated regularly (Lucassen & Schmettow, 2011). However, this open structure makes it very hard, if not impossible, to determine the true source of the information (Lucassen & Schraagen, 2010), which is one of the most frequently used heuristics to evaluate information credibility (Chaiken & Maheswaran, 1994). Furthermore, this also makes it hard to determine whether or not the source is knowledgeable and reliable, which are the two factors that influence trustworthiness (Hovland & Weiss, 1951 – 1952). Therefore, it is necessary to evaluate the credibility of information in other ways.

One of the most important requirements for evaluating credibility is that the user is motivated to evaluate. As is shown in Metzger's dual processing model of web site credibility evaluation (2007), the information will not be evaluated if the user is not motivated to do so. Fogg (2002, 2009) also recognized motivation as an essential factor for behavior. Motivation is affected by the user's perception of the risk of low-quality information. For example, a person that is looking for answers to health questions will most likely perceive a greater risk of trusting false information than a person that is looking up information for their own entertainment.

A second factor that affects credibility evaluations is ability. When a person is motivated to evaluate the information, the way this is done depends on how capable a user is to evaluate. If a person lacks the necessary abilities, the information will be evaluated heuristically; if a person is able to evaluate, they will take a more systematic approach (Metzger, 2007). A more detailed view of how a person's abilities affect the way information is evaluated is provided in the 3S-model, proposed by Lucassen and Schraagen (2011b). The model shows that domain expertise enables a user to evaluate text based on content features, such as correctness and completeness. However, people usually aim to acquire new knowledge instead of looking up things they already know. Therefore, users' domain expertise generally is very limited, which makes it hard to evaluate the content of information.

When evaluation of information content doesn't work, people turn to their information skills to evaluate the credibility of information. According to Lucassen and Schraagen (2011b), information skills enable evaluation of surface features, such as references. However, lay people are generally not trained to make credibility evaluations. According to Walraven, Brand-Gruwel, and Boshuizen (2009), despite the fact that students are aware that they should make credibility evaluations, they rarely do so. Lucassen and Schmettow (2011) suggest that this might be because they don't know how to assess credibility.

In addition to actively evaluate information by content or surface features, people can also use source characteristics to determine the credibility of information. Source characteristics are, according to the 3S-model (Lucassen & Schraagen, 2011b), enabled by previous experiences with the source. This passive way of evaluating information is less time-consuming than the other two strategies, but when the true source of the information is unknown, as is the case with Wikipedia, it carries the risk of people over- or underrelying on the source. For example, when someone blindly trusts all the information from Wikipedia, they are at risk of also trusting some poor information. When someone considers Wikipedia untrustworthy and avoids using it altogether, they risk missing out on high-quality information (Lucassen & Schmettow, 2011; Lucassen & Schraagen, 2012a).

Another factor that plays a role in evaluating credibility is prominence, as explained in the prominence-interpretation theory (Fogg, 2003). According to this theory, the user first notices an element of the information, and then interprets what this means for the credibility of the information. Prominence is affected by three kinds of factors: user characteristics (motivation, ability, experience, and individual differences), context characteristics (the task of the user), and the content of the website. Interpretations are affected by assumptions of the user, such as heuristics or past experiences; the context, for instance, norms and expectations; the user's knowledge of the subject, and the user's goals. As shown earlier, different users use different techniques to evaluate information, based on their knowledge and abilities. This means that it also varies what factors users deem relevant for evaluating credibility, causing prominence and interpretation to vary between users.

In conclusion, evaluating credibility on the Internet can be difficult, especially when a person does not have (sufficient) domain expertise and/or information skills and when the heuristics of evaluating credibility by judging the source of the information are not effective. However, credibility evaluations are necessary. The open structure of the Internet allows

anyone to put information online, which makes it hard to determine the true source of the information. As people often do not know how to evaluate credibility, a possible solution to these issues is to offer them support.

SUPPORTING ONLINE CREDIBILITY EVALUATIONS

In order to help people make credibility evaluations, Lucassen and Schmettow (2011) proposed to help people evaluating credibility by placing a decision support system on the Wikipedia page. A support system can provide the user with an advice about the credibility of the article. While the presence of a support system could cause people to over-rely on the support and thus risking to also trust incorrect advice (Lucassen & Schmettow, 2011), a large advantage of offering support is that it makes people more able to evaluate credibility.

Once the decision to offer support has been made, there are different sorts of support systems that could be used. Lucassen and Schraagen (2012b) differentiate between reputation systems, which use user-generated indications of credibility; and automated decision support systems, which use automated algorithms used to calculate trust values. Jøsang, Ismail, and Boyd (2007) describe reputation as "a collaborative measure of trustworthiness […] based on the referrals or ratings from members in a community". Resnick, Kuwabara, Zeckhauser and Friedman (2000) formulate three requirements for reputation systems: the entity must be long-lived, so that it serves as a base for expectations of future interactions; feedback about current interactions must be captured and distributed; and the feedback must be used to guide trust decisions. A long-lived entity means that the entity that is subject to the feedback is not able to erase its past feedback through, for instance, changing its identity.

An example of a reputation system is eBay's feedback system, in which buyers and sellers can rate each transaction they make to calculate a measure of trustworthiness for each user. The underlying principle is that users with a higher rating are considered more trustworthy than those with lower ratings. An advantage of such a reputation system is that the scores of users are publicly available to other users, who can use them in their decision making process. This corresponds with the second and third requirement of Resnick et al. (2000). Furthermore, users are not able to change their identity and cut the ties with past feedback, which means that the eBay's feedback system also meets the first requirement of Resnick et al. However, a disadvantage is that people are often biased towards rating positively. Resnick and Zeckhauser (2002) found that less than 1% of the ratings on eBay

were negative. In the case of Wikipedia, there is another disadvantage to this approach. As articles are updated regularly, ratings of a previous version of the article are not valid anymore after a revision. This causes an article to have only a small number of ratings before each revision, which violates the first requirement for reputation systems of Resnick et al.

Another way to approach decision support is through automated support, which relies on algorithms to automatically calculate the credibility of the information. Users should be aware of the fact that automated support systems rarely are 100% reliable, and that they therefore should not trust them blindly. According to Parasuraman and Riley (1997), when users over-rely on the support system, their trust is greater than the actual capability of the support, which leads to misuse. When users under-rely, the capability of the system is greater than the users' trust, which leads to disuse.

In the past, various support systems for (websites such as) Wikipedia have been developed. An early example is a tool proposed by McGuinnes, Zeng, Da Silva, Ding, Narayanan, and Bhaowal (2006). This automated support system calculates trustworthiness by the number of times an article is linked to in other articles. However, while a link in another article can be seen as a sign of trustworthiness of the linked article, the true reason why some articles have low link-ratios is hard to determine. Therefore, McGuinnes et al. deem this tool unfit to be a measure for trustworthiness on its own, and they suggest it is rather used alongside other measures.

Another example of an automated support system is WikiTrust. This tool was developed by Adler, Chatterjee, De Alfaro, Faella, Pye, and Raman (2008), and it "computes quantitative values of trust for the text in Wikipedia articles". WikiTrust is based on the assumption that the longer words remain unchanged, the more trustworthy they are. In computing trustworthiness, the reputation of an author's work is considered. Furthermore, the tool shades the background of recently changed words dark orange to indicate that the changes might be untrustworthy. As the words remain unchanged for a longer amount of time, the background color fades to white. Because of the fact that a user's reputation is taken into account when calculating a trust value, the trust is more robust and resistant to tampering. While another advantage of the tool is that it provides users with a clear visual cue of recently edited information, Lucassen and Schraagen (2011a) found that users didn't know how to incorporate that information into their own credibility evaluation process. Another disadvantage is that not all changes that are made are substantive, as is the case with

typographical corrections, but they still cause the changed sections to be highlighted as possibly untrustworthy.

Kittur, Suh and Chi (2008) designed a support system that visualized the past of both the Wikipedia edits as well as contributing authors. Criteria used are percentage of text edited by anonymous users, who made the last edit (an anonymous user or a user that has made many edits), stability of the article as measured by changed words, and a diagram with the edit history of the article. It is up to the users of the support system to interpret the visualized data and to implement it into their credibility evaluation. Initial validation of the support system showed that the tool supports people's credibility evaluations, indicating that presenting users with information about the stability of articles and authors may be an effective approach for a support system.

A tool that has the potential to provide users with similar characteristics about Wikipedia articles as the tool proposed by Kittur, Suh and Chi (2008) is Wikipedia's Article Feedback Tool. Although this tool is not being developed as a decision support system, it has te potential to also serve as such.

WIKIPEDIA'S ARTICLE FEEDBACK TOOL

The Article Feedback Tool (AFT) is a tool that is being developed by Wikipedia and that aims to improve article quality and to get more users to contribute in writing and editing articles (WikiMedia Foundation, 2012). In the version that is currently online – version 4 – users can rate an article on four scales: trustworthiness, objectiveness, completeness and whether or not the article is well-written. From these ratings, an advice about the credibility of the article can be generated. Figure 1 depicts the AFT as it is currently shown beneath each article

While Wikipedia intends to use this advice to encourage more users to participate in the writing- and editing process, it could potentially also be used to help users in their credibility evaluations. In terms of Jøsang et al. (2007), the Article Feedback Tool can be characterized as a 'centralized reputation system', which means that users are able to provide ratings, out of which a central authority computes an overall score. A disadvantage of the AFT is that the article changes with each edit, causing previous ratings to expire, which violates the long-lived entity requirement for reputations systems by Resnick et al. (2000). However, in order to deal with this, Wikipedia decided that instead of letting the ratings

expire with each revision, they expire after 30 revisions. Furthermore, the AFT is capable of collecting and distributing feedback, as is also a requirement for reputation systems. What we will study here is whether or not the AFT meets the third requirement of being used for trust decisions.

.

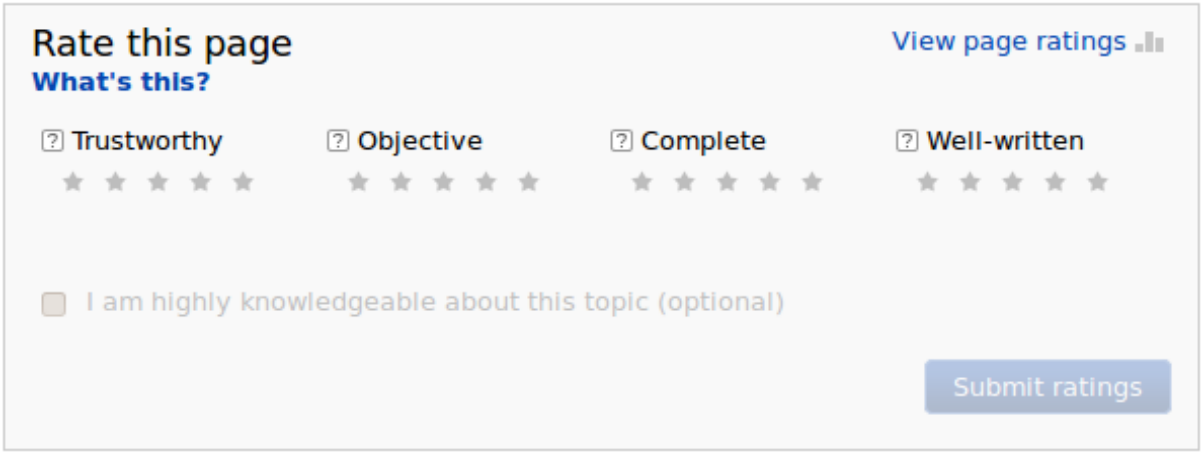

Figure 1. Wikipedia's Article Feedback Tool

In developing the fifth version of the Article Feedback Tool, the WikiMedia Foundation – the foundation behind Wikipedia – is studying the effect of different placements of the feedback button through which users will have access to past feedback (WikiMedia Foundation, 2012). With changing the location of that element on the web page, it is – in correspondence with Fogg's prominence-interpretation theory – possible that the prominence of the element also changes. This could lead to a different impact of the feedback button on the user's credibility evaluation. Therefore, the effect the placement of a support system on a web page has on a person's credibility evaluation will also be studied in this thesis.

HYPOTHESES

The first step in this experiment will be to see whether or not the presence of a support system on a Wikipedia article influences people's credibility evaluations. In order to do that, we first need to define how such an influence would be measurable.

As we defined trust earlier as a person's willingness to depend on something, it became clear that the risk that is involved in trusting can be diminished by performing credibility evaluations. Through these evaluations, a person tries to estimate as accurately as

possible what the information quality is, in order to decide whether or not to trust it. According to Kelton, Fleischmann and Wallace (2008), trust mediates between information quality and the usage of the information, and therefore, information quality could be seen as an indicator for the trustworthiness of the information. Information quality is influenced by several factors, such as accuracy, currency, believability and coverage.

The goal of a support system for credibility evaluations is to help users estimate the information quality as closely as possible to the actual information quality, in order for them to come to an accurate trust judgment. Therefore, the first hypothesis is as follows:

> Hypothesis 1: The perceived trust by users that are provided with support is more in accordance with the actual article quality than the perceived trust by users that are not provided with support.

As an extension of Hypothesis 1, we examine whether or not the placement of the support system is of influence. We suspect that when the support system is placed more prominently on the page, for instance at the top of the page, the influence of the support system will be greater than when it is placed in a less prominent place, for instance at the bottom of the page. In other words, when the support system is placed at a more visually prominent location, users are more likely to consider its advice in their credibility evaluations. Hence, the second hypothesis is as follows:

> Hypothesis 2: The perceived trust by users who are provided with more prominently placed support is more in accordance with the actual article quality than the perceived trust by users who are provided with less prominently placed support. The perceived trust of users who were provided with less prominent support will be more calibrated with the actual article quality than the perceived trust of users that are not presented with support at all.

METHOD

PARTICIPANTS

The experiment was conducted with 38 participants (11 male, 27 female) with an average age of 20.92 (*SD* = 4.89). 22 of the participants had the Dutch nationality, 15 were German and 1 was Hungarian. The participants took part in this experiment as a mandatory part of their studies in behavioral science which required them to take part in several experiments. Participants were able to sign up for the experiment online, and for their participation they were awarded credits they needed to pass the mandatory experiment attendance.

TASK

The participants were shown ten articles from the English Wikipedia. To prevent them from skipping ahead to the questions without reading at least some of the article, participants were asked to answer a question about each of the ten articles they were presented with. The answers to these questions could be found in the articles. In addition, the participants were asked about their trust in the article and their familiarity with the topic it concerned. After the participants had answered all the questions, they could proceed to the next article.

STIMULI

In this experiment, the participants were presented with modified offline versions of 10 Wikipedia articles, which were selected to be of both high and low quality. The high-quality articles were classified as featured articles by the Wikipedia Editorial Team, indicating that they are considered by Wikipedia editors to be among the best articles on Wikipedia. The low-quality articles were classified as start class articles, meaning that the articles were incomplete and lacked reliable resources (WikiMedia Foundation, 2012) The articles were modified to add support in certain conditions, and hyperlinks in the text were made inactive to prevent participants from visiting other websites or articles. However, links leading to the support system were intact to provide the participants with the possibility to view the page ratings it provided. The participants were not informed about viewing offline versions of the articles and the presence or absence of support. Participants were able (but not informed) to use the browser's search function throughout the task. As part of their education is given in

English, we chose to use articles from the English Wikipedia. After the experiment, participants were asked to rate on a 7-point Likert scale if they felt that the fact that the articles were in English impeded them in completing the experiment. Analysis of the familiarity ratings on a scale ranging from 1 (no impediment by the English language) to 7 (impediment) showed an average rating of 3.65 (*SD* = 2.0), indicating that the participants on average experienced some difficulty with the language, but that the language did not pose such a problem that the participants felt as if completing the task was impossible due to it.

PROCEDURE

Upon their arrival, the participants were asked to sign an informed consent, after which they were given instructions for the task they were about to perform. Before the start of this task, they were asked to fill out a few basic demographic questions, as well as questions about their level of experience with and trust in Wikipedia. 26.3% of the participants reported to use Wikipedia every month, 57.9% said to use it every week and 15.8% reported daily use of the encyclopedia. Analysis of the trust ratings on a 7-point Likert scale ranging from 1 (not trustworthy) to 7 (trustworthy) showed an average trust in Wikipedia of 5.11 (*SD* = 1.27), indicating a fairly high level of trust.

After the participants had viewed all ten articles and answered their accompanying questions, they were asked additional questions, such as whether or not they had been aware of the presence of support, if and how they had used the advice the support system gave, and whether they had used the search function. Finally, the participants had the option to leave comments.

DESIGN

The experiment had a 2 (article quality: high/low) x 3 (presence of support: none/prominent/non-prominent) design. Participants were assigned randomly to a condition, and the order of the articles was controlled through Latin square.

ARTICLES

Participants were shown both a high-quality and a low-quality article in each of five categories. Table 1 shows the titles of the articles that were used in each category.

Table 1. Low Quality and High Quality Articles Used in the Different Categories

	Low quality	High quality
Arts	Museum of Fine Arts (Budapest)	Western Chalukya Architecture
Geography	Pyin U Lwin	Yarralumla, Australian Capital Territory
History	Flight to Varennes	Nyon Conference
Mathematics	Euclid's Optics	Polar Coordinate System
Science	B-type main sequence star	Comet Shoemaker-Levy 9

We chose articles of which we expected the participants to have little to no knowledge, because Lucassen and Schraagen (2012b) found that users that are familiar with the topic of information at hand were less influenced by a support system and also had less trust in such a system. Unfamiliar users however, preferred a support system that focuses on surface features, and since the Article Feedback Tool focuses partly on surface features, we aimed to select articles with which the participants were unfamiliar. Participants were asked to rate on a 7-point Likert scale how familiar they were with the topic of each of the articles. A manipulation check was performed after the experiment to check if we really selected articles with which the participants were unfamiliar.

The questions were chosen in such a way that the position of the answer in the text varied slightly between the articles. Although Hensel (2012) found no conclusive evidence that the time people spend searching for an answer is of influence on their trust judgment of Wikipedia articles, we formulated questions of which the answers could be found in the middle or last part of the article to prevent participants from finding the answer at first glance in the opening sentences of the articles. This way, participants had to spend some more time reading the article than just looking at the introduction. To prevent a learning effect, the place in the article where the answer could be found varied between the articles. For example, in one of the articles, the answer to the question could be found around the middle of the article, while in another article the answer was located more towards the end of the article, causing the participants to have to scroll further down to read it. Analysis of the answers participants gave to the questions showed that participants on average answered 73.16 percent of the questions correctly (SD = 18.76). This suggests that they did indeed read the articles, at least up to the point where the answers could be found.

Since the aim of this study is to find out whether or not prominence of a support system is of influence on people's credibility evaluations, we manipulated the prominence of the support system in the experiment. In one condition, the participants were shown the Wikipedia articles without any manipulations. In the second and third condition, the articles had been manipulated to include support. Key difference was that the support system in the non-prominent support condition was found at the very bottom of the page, without any links from elsewhere in the article to direct the participants to the support system. This condition corresponds to the way the Article Feedback Tool is currently placed in Wikipedia articles. In the prominent support condition, we added a link in red color and bold typeface underneath the title of the article, making it highly visible and thus more prominent. The link directed the user to the support system at the bottom of the page when clicked. Clicks on both the Article Feedback Tool and the prominent link on top of the page were recorded.

Figure 2 shows the Article Feedback Tool as users in the prominent and non-prominent support condition encountered it at the bottom of articles. When the link in the top right corner ('View page ratings') was clicked, the tool provided the user with information about previous votes on the article, as shown in Figure 3. It displays the average rating on four factors, as well as the number of people that voted. Figure 4 shows the link to the Article Feedback Tool in the prominent condition.

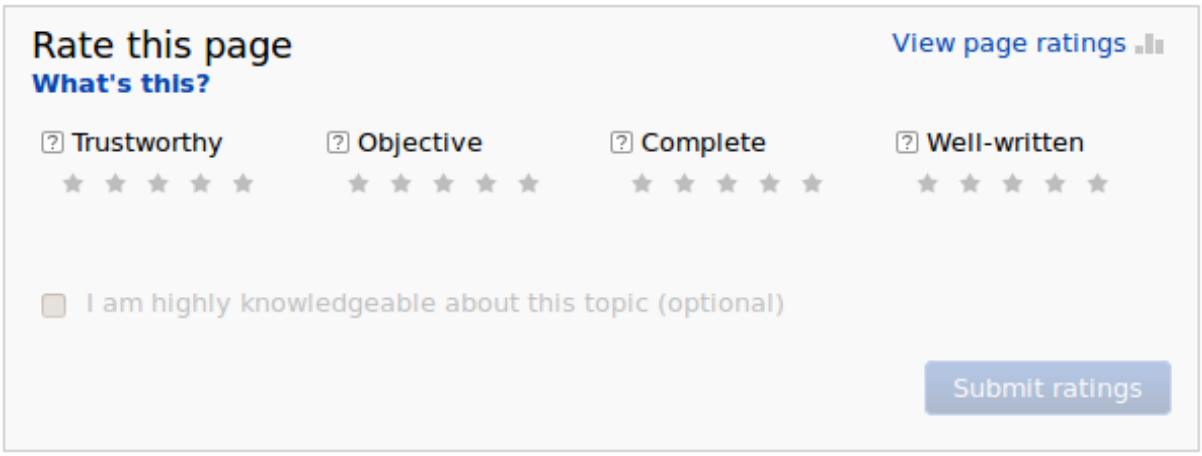

Figure 2. Wikipedia's Article Feedback Tool, as presented to users in the prominent and non-prominent support conditions

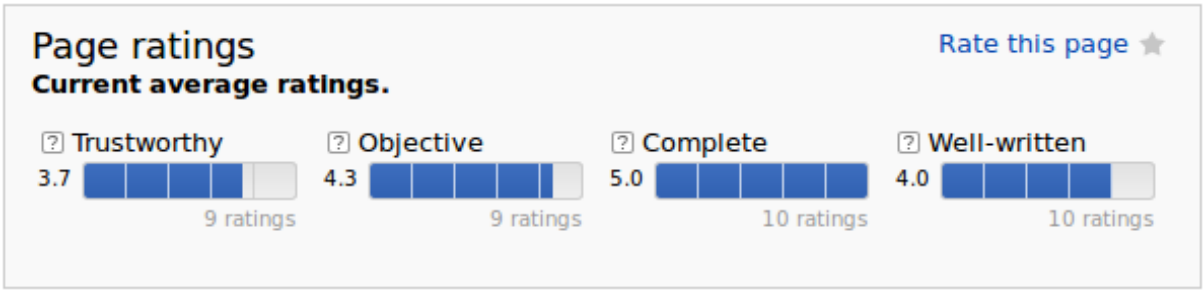

Figure 3. The Article Feedback Tool displaying page ratings

Figure 4: Title with the link to the Article Feedback Tool in the prominent condition

The Article Feedback Tool in the prominent and non-prominent support conditions was manipulated to give advice according to the quality of the articles. To create a clear distinction between the ratings of high-quality and low-quality articles, low-quality articles received ratings between 1.2 and 1.8 out of five, while high-quality articles were given a rating between 4.2 and 4.8.

Between low-quality and high-quality articles, the number of ratings they received varied. To rule out any influence of these numbers, we chose to give all the articles in the prominent and non-prominent condition a representative number of ratings. Out of 50 high-quality Wikipedia articles, we calculated the average number of ratings. Based on this average of 30.36, we manipulated the Article Feedback Tool in the prominent and non-prominent conditions to have between 30 and 40 ratings. In accordance with the real page ratings, ratings varied slightly between the four scales of the Article Feedback Tool.

MEASURES

TRUST IN THE ARTICLE

The participants were asked to rate their trust in each article on a 7-point Likert scale. They were also asked to give a motivation for their rating, which was coded afterwards using a coding scheme Hensel (2012) based on the 3S-model by Lucassen and Schraagen (2011b). Because of the possibility of participants mentioning the support system in their motivations, a fourth 'S' (for 'support') was added to the coding scheme. Furthermore, during the coding of the motivations, a new feature was added to the 'Other' category as some participants mentioned the date the article was last edited as explanation for their trust ratings.

To calculate the inter-rater reliability, 20% of the motivations were coded by a second person. A Cohen's Kappa of 0.79 was found ($p < 0.001$), indicating a substantial agreement.

FAMILIARITY RATINGS

To check our familiarity manipulations, participants were asked to rate on a 7-point Likert scale how familiar they were with the topics of the articles.

DATA ANALYSIS

The hypotheses were tested using a repeated measures ANOVA.

RESULTS

FAMILIARITY MANIPULATION CHECK

We aimed to provide the participants with articles on topics they were unfamiliar with. Analysis of the familiarity ratings showed an average rating of 1.61 (SD = 0.60) on a scale from 1 (unfamiliar) to 7 (familiar). This indicates that the participants were indeed quite unfamiliar with the topics they were presented with.

CLICKS

During the experiment, clicks on the 'View page ratings'-link on the Article Feedback Tool and on the prominent link on top of the page in the prominent support condition were recorded. However, only ten clicks were recorded during the course of the entire experiment, of which one click on the top link and two on the bottom link by a participant in the prominent support condition, and the other seven by a participant in the non-prominent support condition.

TRUST RATINGS

Analysis of trust ratings in the Wikipedia articles showed a significant difference of ratings on high-quality and low-quality articles ($F(1, 34) = 56.42$, $p < 0.001$), but no interaction between trust and support ($F(2, 34) = 0.95$, $p = 0.395$). Table 2 shows the average trust ratings on both high-quality and low-quality articles in each of the three support conditions. Figure 5 contains a visual representation of the values found in Table 2. Neither of the two hypotheses were supported by these findings, and therefore they were rejected.

MOTIVATIONS FOR THE TRUST RATINGS

Analysis of the motivations showed that out of the total amount of 373 motivations, support was only mentioned three times. Post hoc inspection showed that these three motivations all were given by the only participant in the non-prominent support condition that had clicked the Article Feedback Tool. As can be seen in Table 3, participants most often referred to surface features in their motivations, especially to the presence of references and their quality (or lack thereof). After references, participants most often mentioned the length of the article as motivation for their trust ratings.

Table 2. Average trust in high and low-quality articles in the three support conditions. Trust was rated on a 7-point Likert scale ranging from 1 (not trustworthy) to 7 (trustworthy)

Condition	High quality	Low quality
No support	5.68 (*SD* = 1.04)	4.71 (*SD* = 1.02)
Non-prominent support	5.37 (*SD* = 0.85)	4.35 (*SD* = 1.12)
Prominent support	5.89 (*SD* = 0.73)	4.76 (*SD* = 0.80)

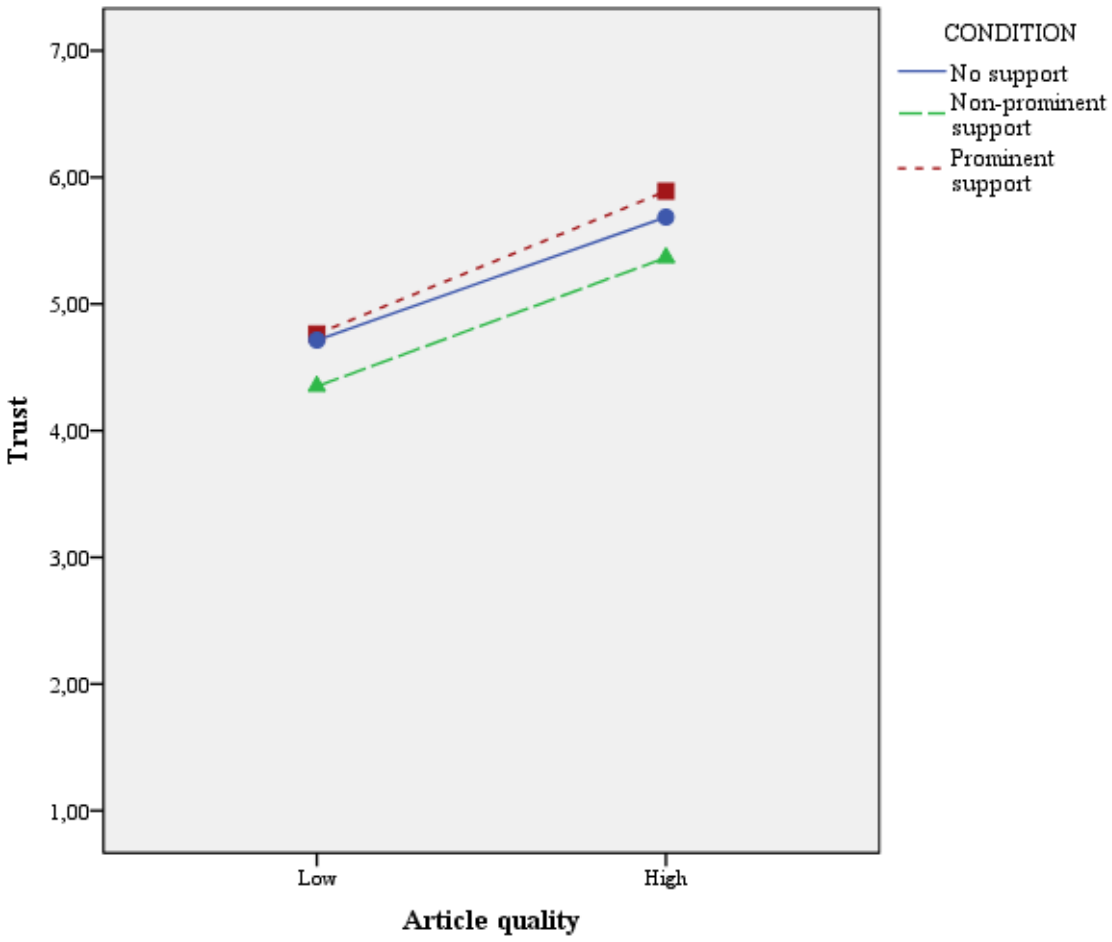

Figure 5. Visual representation of the average trust values from Table 2.

Table 3. Percentages and totals of motivations, based on the coding scheme devised by Hensel (2012)

Category	Percentage of motivations
Semantic features	**14.5% (n = 54)**
Accuracy	4.3% (n = 16)
Completeness	4.6% (n = 17)
No knowledge	4.8% (n = 18)
Objectivity	0.8% (n = 3)
Surface features	**78% (n = 291)**
References	48.5% (n = 181)
Topic	3.2% (n = 12)
Writing style	1.9% (n = 7)
Length	17.2% (n = 64)
Images	0.5% (n = 2)
Structure	2.4% (n = 9)
First impressions	1.9% (n = 7)
Numbers	2.4% (n = 9)
Source features	**2.4% (n = 9)**
Author	2.1% (n = 8)
Wikipedia	0.3% (n = 1)
Support features	
Rating	0.8% (n = 3)
Other features	**4.3% (n = 16)**
Verifiability	0.5% (n = 2)
Edit date	3.5% (n = 13)
Other	0.3% (n = 1)
Total	100% (n = 373)

DISCUSSION

In this paper, we focused on Wikipedia's Article Feedback Tool (AFT) as a possible support system, and specifically whether the AFT is of influence on people's credibility evaluations of Wikipedia articles. The first question this paper aimed to answer was whether the AFT meets the third requirement for reputation systems as formulated by Resnick, Kuwabara, Zeckhauser and Friedman (2000). This requirement states that a reputation system must serve as a guide for trust judgments, or as Resnick et al. put it: "People must pay attention to reputations". We expected that participants that were provided with the Article Feedback Tool would estimate the information quality more in accordance with the actual information quality than participants in the control group, but the data that were gathered showed no interaction effect of support and quality on trust. This finding is supported by the few clicks on the AFT that were recorded during the course of the experiment and by the fact that support was hardly ever mentioned in the participants' trust motivations, indicating that participants generally did not pay attention to the AFT.

The fact that participants didn't seem to notice the AFT indicates that the AFT did not influence people's credibility evaluations, because according to Fogg (2003), for an item on a website to have any influence on these evaluations, it has to be noticed first. A possible explanation for the participants not paying attention to the AFT may be that they simply did not see it. However, the AFT was placed below the references of the articles the participants were shown, and as the questions about each article were positioned below the AFT, participants had to scroll down past it in order to answer the questions and proceed to the next article. Therefore, it seems unlikely that participants didn't notice the tool at all. Another explanation is that they did see the tool, but decided not to use it because they deemed it not valuable towards the process of evaluating credibility, thus lowering the impact of the AFT on credibility evaluations.

An alternative explanation for participants not using the Article Feedback Tool may be that they were not motivated to make credibility evaluations. Both Fogg (2003, 2009) and Metzger (2007) recognize that a person must be motivated to evaluate credibility. It was expected that, in accordance with the 3S-model (Lucassen & Schraagen, 2011b), due to being unable to rely on their knowledge about the topics of the articles and explicitly being asked about their trust in the articles, participants would be more motivated to evaluate surface features, especially the AFT. The data shows that participants were in fact motivated to make

credibility evaluations because they were able to distinguish between high-quality and low-quality articles, and because they indeed mostly focused on surface features. However, it seems that despite this, they were not motivated to use the AFT in their evaluations.

A simple explanation for this finding may be that the AFT did not quite function the same way as other surface features in the article. Instead of the page ratings being immediately visible to the participants, they first had to click a link in the tool to view them. This way, the ratings were not immediately available for use in credibility evaluations, which may explain why participants based their evaluations on more readily available surface features instead of the AFT.

The fact that participants seemed motivated to evaluate credibility but not to use the AFT may be explained using the Dual Processing Model of credibility assessment by Metzger (2007). According to that model, when a person is both motivated and able to assess credibility, information is evaluated systematically. However, when someone is motivated but lacks the ability to evaluate, the evaluation is based on heuristics. It was expected that the latter would be the case with the participants in this experiment, causing them to resort to use the Article Feedback Tool, but the data suggest that the participants were both motivated and able to evaluate credibility of the articles. In fact, all the participants were academic students with trained information skills. Due to their academic background, they may have been so confident in their information skills that they felt no need for a more heuristic approach, such as using the AFT. However, on several occasions participants stated in their trust motivations that they had no knowledge on the topic of the article and that they found it hard to judge credibility.

Since the data suggests that the participants were motivated to evaluate credibility, the fact that they did not use the AFT may have yet another explanation. Perhaps the differences in article quality between high-quality and low-quality articles were so obvious that they felt no need to use a support system in their evaluation process. Post hoc analysis of the articles that were used in the experiment shows that this may indeed have been the case, as high-quality articles on average contained over four times more words than low-quality articles (3859 vs. 900 words for high-quality and low-quality articles, respectively). In addition to that, high-quality articles on average contained over seven times as many references as low-quality articles (52 vs. 7 references for high-quality and low-quality articles, respectively) and over three times as many images (13 vs. 3.8). In accordance with the finding that participants

mostly relied on surface features to evaluate the credibility of the articles, this suggests that the differences between high-quality and low-quality articles were indeed too salient to cause a need for assistance by the AFT. Furthermore, participants that did have difficulty with evaluating the credibility of the articles may not have turned to the AFT because the meaning of the 'page ratings' on which the tool relies was unclear to them. Participants received no information about the AFT or the 'page ratings', and in addition to that, during the experiment, the link on the AFT that normally gives information about the tool (titled 'What's this?') was accidentally disabled along with most of the other links in the articles. This way, participants had no way of knowing what the AFT or the 'page ratings' were, which further decreased the likelihood of the AFT being used in credibilty evaluations.

FUTURE RESEARCH AND PRACTICAL RECOMMENDATIONS

Since the WikiMedia Foundation is still in the process of developing the fifth version of the Article Feedback Tool, this study may be of practical use. Even though the main goals of the AFT are to improve article quality and to involve more users in editing and writing articles, the AFT could potentially be used as a support system. However, since this study shows that the Article Feedback Tool supports credibility evaluations in neither its current form nor in a more visually prominent form, we recommend it is not used in any of these forms. Instead, we suggest that further research should focus on other, possibly more visually prominent placements of the AFT, and that a larger and more diverse participant sample is used.

CONCLUSION

In this study, no proof was found for the hypotheses that the Article Feedback Tool and its prominence influence users' credibility evaluations. However, participants were able to distinguish between high-quality and low-quality information mostly relying on their information skills. This suggests that people that possess information skills may not need a support system to aid them in their decision making process. This may be different in a non-experimental setting where people only have to evaluate the credibility of one article instead of several articles that varied strongly in quality, as they did in this experiment. Future research is needed to study whether or not the AFT does support credibility evaluations in participants with less trained information skills and in situations where only one article or several articles with less variance in quality have to be evaluated.

REFERENCES

Adler, B. T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., & Raman, V. (2008). Assigning trust to wikipedia content. *WikiSym '08: Proceedings of the 4th International Symposium on Wikis* (8–10 September, Porto, Portugal). doi:10.1145/1822258.1822293

Article Feedback (n.d.). Retrieved June 2, 2012 from the WikiMedia Foundation Wiki: http://www.mediawiki.org/wiki/Article_feedback

Article Feedback Tool version 5 feature requirements (n.d.). Retrieved June 15, 2012 from the WikiMedia Foundation Wiki: http://www.mediawiki.org/wiki/Article_feedback/Version_5/Feature_Requirements

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior, 21*(3), 487-508. doi:10.1016/j.chb.2004.10.005

Bronner, F. & Hoog, R. (1983). Non-expert use of a computerized decision aid. *Analysing and Aiding Decicion Processes, 14*, 281 – 299. doi:10.1016/s0166-4115(08)62239-6

Chaiken, S. & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology, 66*(3), 460-473. doi:

Fogg, B. J. & Tseng, H. (1999). The elements of computer credibility. *Proceedings of the SIGCHI conference on Human Factors in computing systems: the CHI is the limit*, 80-87. doi:10.1145/302979.303001

Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. *Proceedings of CHI EA '03: CHI '03 extended abstracts on Human factors in computing systems*, 722-723. doi:10.1145/765891.765951

Fogg, B. J. (2009). A behavior model for persuasive design. *Proceedings of the 4th International Conference on Persuasive Technology*, 1-7. doi:10.1145/1541948.1541999

Hensel, T. (2012). *Impact of duration of the search on trust judgment of Wikipedia articles.* (Unpublished bachelor thesis). University of Twente, Enschede, The Netherlands. Retrieved from http://essay.utwente.nl/61602/.

Hovland, C. I. & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *The Public Opinion Quarterly, 15*(4), 635-650. doi:10.2307/2745952

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. New Haven, CT: Yale University Press.

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems, 43*(2), 618-644. doi:10.1016/j.dss.2005.05.019

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology, 59*(3), 363-374. doi:10.1002/asi.20722

Kittur, A., Suh, B., & Chi, E. H. (2008). Can you ever trust a wiki?: Impacting perceived trustworthiness in wikipedia. *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 477-480. doi:10.1145/1460563.1460639

Lucassen, T. & Schmettow, M. (2011). Improving credibility evaluations on wikipedia. In C. H. Wiering, J.M. Pieters, & H. Boer (Eds.), *Intervention design and evaluation in psychology* (pp. 282-308). University of Twente, Enschede, The Netherlands.

Lucassen, T. & Schraagen, J. M. (2010). Trust in wikipedia: How users trust information from an unknown source. *Proceedings of the 4th workshop on Information credibility,* 19-26. doi:10.1145/1772938.1772944

Lucassen, T., & Schraagen, J. M. (2011a) Evaluating WikiTrust: A trust support tool for Wikipedia. *First Monday, 16*(5).

Lucassen, T. & Schraagen, J. M. (2011b). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology, 62*(7), 1232-1242.

Lucassen, T. & Schraagen, J.M. (2012a). Propensity to trust and the influence of source and medium cues in credibility evaluation. *Journal of Information Science, 38*(6), 564 – 575. doi:10.1177/0165551512459921

Lucassen, T. & Schraagen, J.M. (2012b). The role of topic familiarity in online credibility evaluation support. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 1233 – 1237. doi:10.1177/1071181312561218

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. *Proceedings of the Workshop on Models of Trust for the Web,* 3 – 131.

McKnight, D. H., &d Chervany, N. L. (1996). The Meanings of Trust. *Technical Report MISRC Working Paper Series* 96(04). University of Minnesota, Management Information Systems Research Center.

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. doi:10.1002/asi.20672

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(2), 230-253. doi:10.1518/001872097778543886

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45-48. doi:10.1145/355112.355122

Resnick, P. & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Advances in Applied Microeconomics, 11,* 127 – 157. doi:10.1016/j.compedu.2008.08.003

Self, C. C. (2009). Credibilty. In D. W. Stacks & M. B. Salwen (Eds.), *An integrated approach to communication theory and research* (pp. 435 - 456). New York: Routeledge.

Version 1.0 Editorial team assessment (n.d.). Retrieved January 27, 2013 from Wikipedia: http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Computers & Education, 52*(1), 234-246.