Predicting Trust in Wikipedia Articles

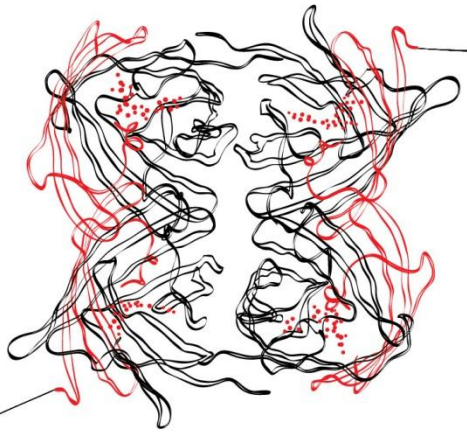
BACHELORTHESIS:

K.L. Cheung

BEGELEIDERS:

T. Lucassen

J.M.C. Schraagen



UNIVERSITEIT TWENTE.

Abstract

Wikipedia is a very popular source of encyclopedic information and studies have shown that its information quality is high. Due to the open source character, evaluation of the source is not possible. Therefore users need to use other features to assess the credibility of the articles. This study focused on measurable features, which are in this case 'surface' features. Studies suggest some important features to predict article quality and that are deemed important by the users themselves. These are 'references', 'internal links', 'pictures', and 'length'. An online experiment was conducted to test whether these features could positively predict the perceived credibility of articles. We did not find any effects of our manipulations. The manipulated features reflected that, based on these features, participants did not trust 'featured articles' more than 'random articles'. However, many motivations of the participants were about these features, especially the number of references. This means that the participants did think the features were important. One explanation for this result is that the interpretation may be the same for both conditions of our manipulations. Models for evaluating computer credibility may explain this phenomenon.

Statement of contents

1. Introduction

As of July 2011, Wikipedia was listed as the seventh most visited website[1]. It delivers online high quality encyclopedic information that is up to date. Wikipedia has much more visitors than sites with similar purposes. Wikipedia is also a topic of a lot of debate because the content of the site can be changed by everybody in the world. Lim (2009) studied why people use Wikipedia and found that college students tend to use Wikipedia when they have positive outcome expectations. This is influenced by information utility. Despite the open-source nature, Giles (2005) has proven that its articles are of high quality. 42 articles on scientific topics were matched to the articles in the Encyclopaedia Britannica. Encyclopaedia Britannica contained three errors on average per article while Wikipedia contained just four. Fallis (2008) argued that Wikipedia's reliability compares favorably to the reliability of traditional encyclopedias. Furthermore Wikipedia has a number of other epistemic virtues such as speed. However, because content on Wikipedia is written by many people and everybody can add information to its articles, there is always the risk of erroneous information. Therefore users need to evaluate the articles on their credibility to decide how much they trust the articles.

When discussing trust it is important to give a definition for the concept. There is a difference between trust and credibility. Credibility can be described as perceived information quality while trust also involves the notion of willingness to depend on the credibility of information. This means that trust involves a certain risk that a user takes by using the information (Kelton, Fleischmann and Wallace, 2008). This paper uses the definition of Kelton et al. (2008) of the concept 'trust': the mediating variable between information quality and information usage.

It may be difficult to evaluate an article on its credibility (Lucassen and Schraagen, in preparation). Support tools have been developed to support users in their evaluation of credibility. An example of such a tool is WikiTrust by Adler, Chatterjee, de Alfaro, Faella, Pye and Raman (2008). This tool colors the background of each word in an article. The shade represents credibility ranging

[1] http://www.alexa.com/topsites

from low credibility (dark orange) to high credibility (white) The colors are based on the age of a contribution and the reputation of the author. The former is measured by the number of edits to the article in which the word survived the editing. The latter is measured by the survival rate of the author's contributions.

A potential complication with such tools is that the user might use other criteria than the tools focus on to assess an article's credibility. If concepts the user think are important themselves do not match with concepts from the support tool, the user may not accept the system or use it ineffectively. Therefore it is important to know what criteria users use to evaluate an article's credibility. In this study we investigate whether surface features of articles, such as 'references', 'internal links', 'pictures', and 'length', positively influence trust judgments. We will now discuss the strategies users may use to evaluate credibility of information.

1.1 3S-model

The 3S-model proposed by Lucassen and Schraagen (2011) makes a distinction in two types of user expertise on which trust judgments depend, namely domain expertise and information skills. A third strategy to form trust judgments is source experience. By following the strategy of domain expertise, users assess semantic aspects of the information, such as accuracy and neutrality. Users with this strategy use prior knowledge on the topic to form a trust judgment. By following the strategy of information skills, users assess surface features, such as the length of an article and the number of references. Instead of assessing semantic or surface features, the user may follow the strategy of source experience. By following this strategy the user relies on earlier experiences with the source of the information. Users with varying source experience, domain expertise, and information skills may vary in their perceived credibility of the same article. Lucassen and Schraagen (2011) validated the 3S-model and found that domain experts were indeed influenced by the accuracy of the presented information. Users differ in what features to base their trust ratings on (semantic, surface and source). In this study, we investigate the influence of some surface features on trust. Therefore, we

controlled for source and expertise in our articles, the other strategies users can employ to base their trust judgments on.

1.2 Trust and surface features

Lucassen and Schraagen (2010) were interested in features used by users of Wikipedia in their assessment of credibility. They asked participants to perform the Wikipedia Screening Task. This task involves rating the credibility of Wikipedia articles presented in the experiment, without specifying how. In the study by Lucassen and Schraagen (2010), participants performed this task while thinking aloud (Ericsson and Simon, 1984). Participants were presented ten Wikipedia articles which were slightly manipulated to remove any obvious cues of credibility or quality. An example of such an obvious cue is an info box which signals flaws in the quality of an article. Participants were asked to rate the credibility on a 7-point Likert scale. Then the utterances were extracted from the think aloud procedure and classified. Categories mentioned most often were textual features, references and pictures. Also the introduction and internal links were mentioned often.

With this information, we can speculate when an article meets the criteria to be trusted by the users. Trust can be seen as an assessment of information quality which involves uncertainty. One study on the features of Wikipedia articles on information quality was carried out by Stvilia, Twidale, Smitch and Gasser (2005). They proposed seven Information Quality metrics which can be evaluated automatically. These metrics were successful in indirectly assessing Wikipedia article quality corresponding to the set of criteria of 'featured' articles. 'Featured' articles are articles that survived a rigorous nomination and peer-review process that only one article in every thousand makes the cut. The difference between these studies and our study is that these studies focus on quality, while our study investigates on the 'perceived' credibility.

In this study, trust in Wikipedia articles means that users perceive the articles to be dependable. In order to form trust judgments of Wikipedia articles, users have to evaluate the credibility of the articles. In a study by Lucassen and Schraagen (2010) it was found that college

students rate high quality articles as more credible than low quality articles. This means that college students were successful in evaluating credibility of the articles. Based on the criteria by the Wikipedia Editorial Team, which assess the quality of articles on Wikipedia manually[2], the mentioned features in the study by Lucassen and Schraagen (2010) are important for the quality of articles. If the mentioned features are considered important for high quality articles, then these features may have deemed important for users as well when evaluating credibility to assess their trust in the articles.

1.3 Hypotheses

We manipulate Wikipedia articles on several 'surface' features which may be important for users when evaluating credibility of articles to form trust judgments. We investigate the predictive value of each of the mentioned measurable features on the assessment of credibility by users of Wikipedia. These features are: 1. references (number of references), 2. internal links (number of internal links), 3. pictures (number of pictures), and 4. length (number of words. In a study by Lucassen and Schraagen (2010) was found that college students rate high quality articles as more credible than low quality articles. Therefore the difference in quantity of these 'surface' features between 'featured' articles and 'random' articles may influence the assessment of trust. We test whether users trust 'featured' articles more than 'random' articles, based on 'references', 'internal links', 'pictures', and 'length'.

In the 'text' the category, 'references' is mentioned very often. This leads to the following hypothesis.

Hypothesis 1. *Increases in references of Wikipedia articles have a positive influence on trust.* One of the features most often mentioned by participants in the experiment by Lucassen and Schraagen (2010) was the number of 'references' (9.59% of the utterances). Therefore we highly suspect that the amount of references influences trust. Dondio, Barrett and Weber (2006) found that

[2] http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

'featured' articles are referenced best. Sometimes 'featured' articles lose their 'featured' status when the article doesn't hold on the criteria anymore. One of the most identified reason for evoking 'featured' article status is 'verifiability' which means the absence of references in articles. This indicates that 'references' are deemed important for Wikipedia users for assessing credibility.

The feature category 'pictures' is also one of the most mentioned cues. Quality and quantity of that category are mostly mentioned but quality is very subjective and hard to measure while quantity (number of pictures) is measurable. Therefore we include the quantity of pictures in our research which leads to the following hypothesis.

Hypothesis 2. *Increases in pictures of Wikipedia articles have a positive influence on trust.*

The number of 'pictures' is quite often mentioned (12.55% of the utterances). Therefore we suspect that the number of pictures influences trust. In a study by Stvilia et al. (2005) it is noted that the median of the number of pictures in 'featured' articles is five while the number of images in 'random' articles is zero. This means that high quality articles on Wikipedia have pictures while most lower quality articles don't have any images. Steinbrück, Schaumburg, Duda and Krüger (2002) found in their study that embedding a photograph of a company's representative will increase the perceived credibility of an online-vendor. This is a different context and depicts a specific sort of image. However, the idea of including images in articles to increase trust may be applicable to Wikipedia articles as well. After all, 'featured' articles do include them while most 'random' articles do not.

Internal links is mentioned regularly and is measurable quantitatively. Therefore we test the following hypothesis.

Hypothesis 3. *Increases in internal links of Wikipedia articles have a positive influence on trust.*

Participants in the study by Lucassen and Schraagen (2010) also mentioned internal links as a feature to base their assessment of credibility on (1.48% of the utterances). According to the study by Stvilia et al. (2005) the median of number of 'internal links' in 'featured' articles is 206 while the median in 'random' articles is just 17. This indicates that high quality articles have many more 'internal links'

than lower quality articles. If high quality articles have more 'internal links' than 'random' articles, then 'internal links' may positively influence trust. In an experiment by Mcguinness, Zeng, Silva, Ding, Narayanan, and Bhaowal (2006) it was found that 'featured' articles have the highest link-ratio values while 'clean-up' articles have the lowest. 'Clean-up' articles are articles that are below the quality standard of Wikipedia and are typically manually marked by Wikipedia administrators or other authors. The finding means that relatively more articles link to 'featured articles' than other articles. We are interested if articles will also be perceived more credible when the number of internal links increases.

The 'length' feature in the 'text' category is mentioned quite often by the participants. This is measurable because 'length' can be expressed by the number of words. The users may think that if a piece of text is long then someone put in the effort of writing it and that influences trust (Palfrey and Gasser, 2008). This leads to the following hypothesis.

Hypothesis 4. *Increases in length of Wikipedia articles have a positive influence on trust.* With this hypothesis we want to examine if the feature 'length' has a positive influence on trust ratings of articles. In the experiment by Lucassen and Schraagen (2010) 'length' was mentioned regularly (3.31% of the utterances). We suspect that 'length' is an important feature in predicting trust. Blumenstock (2008) studied the link between length and the quality of Wikipedia articles. He made the distinction between high quality and low quality articles where high quality were Wikipedia articles that are 'featured' and low quality articles were ' random'. He found that a good predictor of 'featured' and 'random' articles is the number of words.

Besides trust ratings, we are interested in the motivation of its rating for more insights in the strategies used for trust judgments. In the next chapter, we discuss the method used to test our hypotheses.

2. Method

2.1 Participants

We recruited 259 participants (50.2% were male) for our online experiment on different kinds of forums. A short description of the experiment was given and forum users were asked to become participants. When forum users decided to participate they could click on the link and the experiment started. Participants were recruited from different nationalities, which are depicted in Table 1. The average age was 24.9 years (SD=11.1).

Table 1. **Percentage and frequency for each country**

Country	Percentage	*n*	Country	Percentage	*n*	Country	Percentage	*n*
Australia	1.2%	3	**Estonia**	.4%	1	**Russia**	.4%	1
Austria	.4%	1	**Finland**	1.2%	3	**Saudi Arabia**	.4%	1
Belgium	1.9%	5	**Germany**	1.5%	4	**Scotland**	.4%	1
Brazil	.4%	1	**Greece**	1.5%	4	**Singapore**	2.3%	6
Canada	.4%	1	**Hong Kong**	.4%	1	**Slovenia**	.8%	2
Croatia	10%	26	**Ireland**	.8%	2	**South Africa**	.8%	2
Curacao	.4%	1	**Italy**	.4%	1	**The Netherlands**	57.9%	150
Czech Republic	.4%	1	**Romania**	1.5%	4	**United Kingdom**	4.6%	12
						United States of America	9.7%	25

2.2 Task and procedure

We performed an online experiment in which participants were asked to perform the 'Wikipedia Screening Task' (Lucassen and Schraagen, 2010). In this task, participants are asked to rate the credibility of an article, without specifying how. In our experiment, participants were presented with four Wikipedia articles. These articles were actually manipulated versions of the original articles in the form of full-page screenshots[3] of the manipulated articles. Participants were not allowed to visit other websites. Therefore screenshots were presented, which refrained participants from clicking on internal links. However, we had no control if some participants checked the articles online. The

[3] Using Webshot, http://www.websitescreenshots.com

articles were taken from the English Wikipedia as we assumed that the participants had no difficulty reading the articles. First, an explanation of the experiment was provided and when participants were ready they could click on the 'Next' button which led them to some general and demographic questions which were age, gender, country, usage and general trust in Wikipedia. When clicking 'Next' the first Wikipedia article was presented which was about Armilaria Gallica. When participants read the article they had to answer how much they trusted the article on a 7-point Likert scale. They were asked to write down a motivation for their answer as well. Moreover, the participants were asked to rate how familiar they were with the topic on a 7-point Likert scale. Then participants had to click 'Next' to go to the next article. The second article was about Ceres, the dwarf planet, the third article was about PNC park and the fourth article was about Verbascum thapsus. Participants were not able to go back after they clicked the 'Next' button. IP-addresses were registered and a cookie was saved on the participant's computer to ensure that there were no multiple participations by the same user.

2.3 Design

In the experiment, eight conditions with 4 articles each were presented. Every participant gave a trust rating for each article. Therefore we had 1036 articles rated. Articles were manipulated such that they showed either a high amount or low amount of one specific feature. The participants were randomly assigned to one of the eight conditions. Every participant was assigned to the condition with the least participants. Every condition consisted of four articles in which the orders of topics were the same for all conditions. Each of the four articles was manipulated on one specific feature and the order of the manipulations differed between conditions.

‘Featured’ articles were used for our manipulations. The reason is that ‘featured’ articles already had high numbers of features. This way, minimal features needed to be added for the manipulations, which led to a more natural look of the articles. We selected articles which had a high amount of 'references', 'internal links', 'pictures', and 'length'. Then we manipulated the articles on

the four features. Blumenstock (2008) found that article length can discriminate between high and low quality articles with a word count threshold at 2,000 words. By classifying articles with more than 2,000 words as 'featured' and those with fewer words as 'random' it was found that 'length' is a good predictor of quality articles with 96.31% accuracy. Therefore the manipulations were based on this threshold. In a manipulation of one specific feature, all other features were held constant, meaning the middle amount of the features. For 'length' this was 1,800 - 2,200 words and for 'pictures' this was one image. For a realistic appearance, it was necessary to calculate the mean of number of references and internal links for articles with 1,800 - 2,200 words. The mean of the number of references in articles with 1,800 - 2,200 words of all 'featured' articles was 38. For the number of internal links, the mean was 89. The articles were manipulated by deleting the features according to Table 2. In some cases adding internal links and references was necessary because by manipulating 'length', the number of internal links and references were deleted as well. How every feature was manipulated is discussed in depth next and depicted in Table 2.

Table 2. **Eight experimental manipulations for each of the four articles**

	'References' manipulation		'Internal links' manipulation		'Pictures' manipulation		'Length' manipulation	
	1 (high)	2 (low)	3 (high)	4 (low)	5 (high)	6 (low)	7 (high)	8 (low)
Length	1,800/2,200	1,800/2,200	1,800/2,200	1,800/2,200	1,800/2,200	1,800/2,200	2,500/3,000	1,000/1,500
References	53	27	38	38	38	38	53	27
Internal Links	89	89	124	66	89	89	124	66
Pictures	1	1	1	1	3 or more	0	1	1

2.4 Independent variables

2.4.1 'References' manipulation.

This feature was manipulated by presenting articles with many references and articles with few references. Based on Ad-hoc analysis, we classified 53 as a high number of references while 27 was classified as a low number of references in articles.

2.4.2 'Pictures' manipulation.

This feature was manipulated by presenting articles with no pictures and articles with three or more pictures. While articles low on 'pictures' contained no pictures at all, articles high on 'pictures' contained three or more pictures in an article. These numbers were based on the number of pictures in respectively ‘random’ and ‘featured’ articles.

2.4.3 'Internal links' manipulation

This feature was manipulated by presenting an article with many internal links and one article with few internal links. Based on Ad-hoc analysis, we classified 124 as a high number of internal links while 66 is classified as a low number of internal links in articles.

2.4.4 'Length' manipulation.

Based on a study by Blumenstock (2008), this feature was manipulated by creating articles containing more than 2.500 words (high number of words) and articles with less than 1,500 words (low number of words). As an exception, this manipulation had different amount of features to control for the 'references' and 'internal links'. Longer articles also contain more 'references' and 'internal links'. Therefore articles 'low' on 'length' were manipulated with 27 references and 66 internal links. Articles 'high' on 'length' were manipulated with 53 references and 124 internal links. One image was depicted in both conditions. These numbers were based on Ad-hoc analysis.

2.5 Dependent variables

2.5.1 Trust.

Participants were asked to rate how much they trust each article on a 7-point Likert scale. This scale was quite detailed which made it possible to analyze effects of the manipulations. It was possible to test whether the manipulations predicted the trust ratings (positively).

2.5.2 Motivations for the trust ratings.

Participants were asked to write down on what aspects their ratings were based. Filling in a motivation was optional. Motivations were used to gain insight in the strategies used to form trust judgments. We categorized these motivations into each of the strategies proposed in the 3S-model and analyzed whether the surface features we manipulated had a positive influence on trust. We categorized most motivations into one of the three strategies. When motivations did not fit into one of the three strategies they were categorized as 'other motivations'. Two experimenters both analyzed 10% of the same data. Cohen's Kappa was calculated with the resulting value of .950 which indicates a near-perfect agreement.

2.5.3 Familiarity.

In a study by Lucassen and Schraagen (2011), about the role of expertise in trust, it was shown that experts trust articles differently than novices. To make sure that familiarity did not influence the study, general topics were selected in which we did not expect participants to differ in their familiarity. Participants were asked to rate their familiarity with the topic on a 7-point Likert scale for each article. This way we were able to test whether familiarity differed between articles.

3. Results

3.1 Familiarity

A Friedman test was conducted to evaluate whether differences were found in familiarity for the article about Armilaria Gallica (*M*=1.59), for the article about Ceres (*M*=2.24), for the article about PNC Park (*M*=1.68), and for the article about Verbascum thapsus (*M*=1.75). The test was significant χ_2 (3, *N* = 261) = 16.56, *p* < .01. Ceres seems to be rated higher on familiarity than other articles. Knowledge about astronomy may have caused participants to feel that the article is familiar with what they know. Some participants noted that their prior knowledge about astronomy correlates with the article about Ceres. An example is: ''I know a fair bit about astronomy and everything correlates with what I know about Ceres''. Taking out the article about Ceres resulted in no significant influences of the manipulation on trust ratings (*p* > .05).

3.2 Trust

Table 3 shows the coefficients for each feature and whether the feature was significant in predicting trust ratings. Based on the analysis, no significant results were found. 'References', 'internal links', 'pictures', and 'length' were not predictive of trust ratings (p > .05).

Table 3. **Multiple linear regression of the outcome variable trust on the four predictor variables**

Predictor Variable	B	t(1036)	*p*
References	.000	.019	.49
Internal Links	.001	.304	.38
Pictures	.006	.106	.46
Length	.000	.702	.24

Note: one-sided probability values

3.3 Motivations

Most trust ratings were accompanied with motivations (69%). Some motivations were categorized as 'other motivations' (13% of all motivations) as they did not fit in one of the three strategies of the 3S-model. Motivations in which it was indicated that their trust was based on Wikipedia in general were classified as ‘source’ features (12% of all motivations). An example is: ''I learn in my study (journalism) you can't trust an open system like Wikipedia''. Motivations about the content of the article or about pre-existing knowledge of the participant were categorized as ’semantic’ features (14% of all motivations). An example is: '' Things I know are stated correctly''. Motivations about the presentation of the article were categorized as ‘surface’ features (61% of all motivations). An example is: '' I have good trust in the article, because of the references''.

Motivations which were categorized as ’semantic’ features were subcategorized as 'accuracy', 'completeness', 'scope', and 'other'. Motivations indicating the correctness of the information and knowledge of the topic were categorized as ‘accuracy’. An example is: ''As this is a topic that I am somewhat familiar with, I found most of the material in the article to be trustworthy''. Motivations about the depth of the content were categorized as ‘completeness’. An example is: ''It gives a lot of detail on aspects such as the opening and pricing''. Motivations about the range of the information were categorized as ‘scope’. An example is: ''It's not just a plain text as it clearly covers a diverse range of information on the Armillaria Gallic''. Motivations indicating ‘semantic’ features which did not fit in one of these subcategories were categorized as ‘other’.

Motivations which were categorized as ‘surface’ features gave insight in the features used to form trust judgment. Therefore these motivations were categorized as 'references', 'internal links', 'pictures', and 'length'. Two other features which were noted regularly were 'writing style' and overall 'appearance'. Motivations about the readability and the way it was written were categorized as ‘writing style’. An example is: ''Clearly written and well explained''. Motivations about the overall look of the article were categorized as ‘appearance’. An example is: ''Inherent distrust from first appearance, although it appears well done''. The motivations about ‘surface’ features which did not

fit in these subcategories were categorized as 'other'. The percentages for the three strategies and the percentages for the subcategories are depicted in Table 4.

Table 4. **Percentages of the three strategies of the 3S-model**

'Semantic' features (13.9%, n = 99)	
	Accuracy (10.4%, n = 74)
	Completeness (.7%, n = 5)
	Scope (.3%, n = 2)
	Other (2.5%, n = 18)
'Surface' features (61.4%, n = 437)	
	References (37.9%, n = 270)
	Internal links (1.3%, n = 9)
	Pictures (.3%, n = 2)
	Length (5.1%, n = 36)
	Writing style (4.5%, n = 32)
	Appearance (7.7%, n = 55)
	Other (4.6%, n = 33)
'Source' features (11.8%, n = 84)	
Other motivations (12.9%, n = 92)	

4. Discussion

4.1 Influence of surface features on trust

We did not find any influences of 'references', 'internal links', 'pictures', and 'length' on trust. Therefore all hypotheses were rejected. One explanation is that the amount of features in our manipulations did not differ in influence on trust between the conditions, which reflected the amount of features of 'featured' compared to 'random' articles. Lucassen, Noordzij and Schraagen (2011) found that the length of the reference list influenced trust for their participants. However, in their experiment, the short condition contained only five references and the long condition 25. Our manipulation of 'references' contained many more references to reflect realistic numbers.

Interesting is that many participants did note 'surface' features (61% of all motivations). 'References', 'internal links', and 'length' were quite often noted by participants. However, 'pictures' was just noted twice. Besides our manipulated features, 'writing style' and overall 'appearance' were quite often noted as well. The most often noted feature was 'references' with 38% of all motivations. Perhaps participants did think these features were important, especially 'references', but found it hard to interpret them.

4.2 Prominence-Interpretation Theory

One theory that can explain our results is the Prominence-Interpretation Theory (Fogg, 2003). This theory posits that two components are necessary for credibility evaluation, which are 'Prominence' and 'Interpretation'. 'Prominence' is the likelihood that an element is noticed when people evaluate. 'Interpretation' is what value or meaning people assign to an element. 'Prominence' as well as 'Interpretation' are needed when people assess credibility which leads to trust.

'Length', 'internal links', and especially 'references' were noted by many participants which means that the features were *Prominent*. However, the interpretations of the participants of the

number of features could be the same for both conditions in our experiment. For example, participants may have interpreted the number of references in both conditions as high. According to Fogg (2003), the *Interpretation* component is the user's evaluation of a Web site element. This evaluation can either be positive or negative, which forms the impact an element has on credibility assessment. This may mean that in our study, the evaluation of the participants of the features were positive, in both conditions. This indicates that, based on the amount of 'length', 'internal links', and 'references', no difference in trust of participants existed between 'featured' and 'random' articles. This explains the lack of effect found in our study despite all the motivations noting the features, especially 'references'. Many participants in the 'low' on 'references' noted that there were a lot of references. An example is: '' There are a fair amount of references so I assume they back up the given information'', while this participant was in the 'low' condition of 'references'. 'Pictures' was almost never noted. Perhaps the absence of images did not influence trust because this led to the lack of prominence, which was needed for credibility evaluation. The absence of motivations about 'pictures' indicated that the participants in our experiment did not think the number of pictures were important as well.

4.3 'Surface' features and models for evaluating computer credibility

Fogg and Tseng (1999) proposed three prototypical models for evaluating computer credibility. These models can explain the way the participants interpreted the features in our experiment. First, binary evaluation is a strategy for evaluating credibility in which users perceive the product as either credible or not credible without a middle ground. Second, threshold evaluation strategy includes upper and lower thresholds for credibility assessments. Between the two thresholds the perceiver may somewhat trust the product. Last, spectral evaluation strategy is the strategy in which trust gradually increases with product credibility.

Participants may have evaluated different elements in the articles with these strategies to form a trust judgment for the whole article. In that case, it is conceivable that the 'surface' features

in our study were evaluated with either the binary evaluation, or the threshold evaluation strategy by the participants. This is because spectral evaluation is facilitated only when there is high interest in the issue, high ability to process the information, high familiarity with the subject matter, and considerable opportunity to compare various sources. Participants in our experiment had probably not much interest and familiarity with the topics. This may also explain why we found no effects of our manipulations. The lack of effects between the conditions of the manipulations indicates that the participants did not use the spectral evaluation strategy. Participants may have deemed each feature credible by evaluating whether the feature exceeded a threshold (binary evaluation), or the upper threshold (threshold evaluation). Both of our conditions of the manipulations may have exceeded this threshold which led to the same positive credibility impact for both conditions. We noticed that many motivations about our features were positive, in both conditions. An example is: ''The article has many citations and is well written with links to many related topics'', while this participant was in the 'low' on references condition. This may mean that the chosen numbers of the features for the 'low' conditions were considered credible. These numbers were 1,000 – 1,500 words, 27 references, and 66 internal links.

4.4 Motivations and the 3S-model

The 3S-model was used to predict that the participants mainly assessed 'surface' features for trust judgments, while 'semantic' features and 'source' features were kept constant. Our study was similar to the study by Lucassen and Schraagen (2011) which focused on the manipulation of 'semantic' features. They found that, in contrast to participants who used 'domain expertise' (experts), participants who used 'information skills' (novices) were not influenced in their trust judgments by factual accuracy. Our study focused on the manipulation of 'surface' features and its influence on trust. Moreover, we found additional support for the 3S-model. Many motivations (61%) were categorized as 'surface' features. This percentage was higher than in the study by Lucassen and Schraagen (2011) in the novices condition (48%). One explanation is that their study manipulated on

'semantic' features while our study manipulated on 'surface' features. The three categories of the 3S-model covered in both studies about 88% of the motivations. However, the percentage of the motivations of novices which were categorized as 'semantic' features (7%) in the experiment by Lucassen and Schraagen (2011) was half of ours (14%). One explanation is that in their study, 'semantic' features were manipulated. Therefore they compared experts with novices on selected specific topics for the experiment. In our study, general topics were selected to keep familiarity with the topics constant. Novices in their study, selected on their interest, may have been less familiar with the topics than participants in our study. This may have led to more usage of the 'semantic' features in our study, compared to the novices condition in their study. Contrary, in the study by Lucassen and Schraagen (2011), more motivations (of novices) were categorized as 'source' features (34%) than in our study (12%). One explanation is that in the study by Lucassen and Schraagen (2011), each participant viewed only one article. In our study, each participant viewed four articles which made comparison possible. Comparing the articles, which were all from Wikipedia, may have led to less usage of the source when judging trust. The idea is that comparing articles from the same source, based on 'source experience', was not useful for distinguishing the articles. Therefore, participants may have assessed the source more in the experiment by Lucassen and Schraagen (2011) in which comparison was not possible.

4.5 Limitations

In this study we found no effects of the manipulations on trust. Our manipulations were probably not extreme enough to detect any effects.

Participants in the experiment were recruited from online forums. High numbers of participants were needed and therefore recruiting from online forums is a great strategy. This strategy makes it possible to recruit a high number of participants all over the world. However, accountability is low. We had no control on what they were doing during our experiment. If participants compared the manipulated articles to the original ones online, we would not know.

The motivations for trust judgments of the participants were provided optionally, and not all participants did so. This means that our results may not be representative for the entire sample in the experiment.

4.6 Future research

Lucassen and Schraagen (2011) proposed the 3S-model and performed an experiment manipulating ‘semantic’ features. We focused on manipulating 'surface' features, in which we found no effects. The three proposed models for evaluating computer credibility (Fogg and Tseng, 1999) explained our results. Studies should therefore further investigate the relation of the numbers of ‘semantic’ features and these models. Future studies can test which model, the binary evaluation model or the 'threshold evaluation model, is used for each feature. Our experiment can serve as a starting point for further research on trust in Wikipedia and 'surface' features.

4.7 Conclusion

This study has provided new insights concerning credibility evaluation, based on ‘references’, ‘internal links’, ‘pictures’, and ‘length’. We found that the difference in the amount of the features between ‘featured’ and ‘random’ articles had the same credibility impact. Participants may have interpreted the amount of features positive in both conditions of our manipulations. The three models for evaluating computer credibility by Fogg and Tseng (1999) can explain the results. Participants may have used binary or threshold evaluation strategy to interpret the features in our study. This explains why the manipulations had no effect on trust. Furthermore, the motivations provided additional support for the 3S-model.

References

Adler, B.T., Chatterjee, K., Alfaro, L., de, Faella, M, Pye, I., & Raman, V. (2008). Assigning trust to Wikipedia content. In *WikiSym 2008: International Symposium on Wikis.*

Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on Wikipedia. *Proceedings of the 17th International Conference on the World Wide Web*, ACM Press, 1095-1096.

Chevalier, F., Huot, S., and Fekete, J.-D. (2010). WikipediaViz: Conveying article quality for casual Wikipedia readers. *Pacific Visualization Symposium*, IEEE, 49-56.

Dondio, P., Barrett. S., and Weber, S. (2006). Calculating the trustworthiness of Wikipedia articles using DANTE methodology. *IADIS International Conference e-Society.*

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: Bradford Books/MIT Press.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(10):1662-1674.

Fogg, B. J. (2003). Prominence-interpretation theory: explaining how people assess credibility online. In *CHI '03 extended abstracts on Human factors in computing systems*, CHI EA '03, pages 722-723, New York, NY, USA. ACM.

Fogg, B. J. and Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, CHI '99, pages 80-87, New York, NY, USA. ACM.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature*, 438(7070).

Kelton. K., Fleischmann.K.R., Wallace & W.A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology* , *59*.

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.

Lucassen, T. and Schraagen, J. M. (2010). Trust in Wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th Workshop on information Credibility*. WICOW '10. ACM, New York, NY, 19-26.

Lucassen, T. and Schraagen, J. M. (2011). Introducing domain expertise in models of information trust. *Journal of the American Society for Information Science and Technology,* page n/a.

Lucassen, T. and Schraagen, J. M. (2011). 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., Noordzij, M. L., and Schraagen, J. M. (2011). Reference blindness: The influence of references on trust in Wikipedia. In *ACM WebScience 2011*.

McGuinness, D.L., Zeng, H., Pinheiro da Silva, P., Ding, L., Narayanan, D., Bhaowal, M. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. In *Proceedings of the Workshop on Models of Trust for the Web.*

Palfrey, J. & Gasser, U. (2008). *Born Digital: Understanding the First Generation of Digital Natives* (1st ed.). New York: Basic Books.

Steinbrück, U., Schaumburg, H., Duda, S., and Krüger, T. (2002). A picture says more than a thousand words: photographs as trust builders in e-commerce websites. In *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*, pages 748-749, New York, NY, USA. ACM.

Stvilia, B., Twidale, M.B., Smith, L.C., and Gasser, L. (2005). *Assessing information quality of a community-based encyclopedia*. In *Proceeding International Conference on Information Quality*, 442-454.

Sutter, M. and Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior*, 59(2):364-382.