Trust in Wikipedia: How Users Trust Information from an Unknown Source

Teun Lucassen
University of Twente
P.O. Box 215, 7500 AE, Enschede
The Netherlands
t.lucassen@gw.utwente.nl

Jan Maarten Schraagen
University of Twente
P.O. Box 215, 7500 AE, Enschede
The Netherlands
j.m.c.schraagen@gw.utwente.nl

ABSTRACT

The use of Wikipedia as an information source is becoming increasingly popular. Several studies have shown that its information quality is high. Normally, when considering information trust, the source of information is an important factor. However, because of the open-source nature of Wikipedia articles, their sources remain mostly unknown. This means that other features need to be used to assess the trustworthiness of the articles. We describe article features - such as images and references - which lay Wikipedia readers use to estimate trustworthiness. The quality and the topics of the articles are manipulated in an experiment to reproduce the varying quality on Wikipedia and the familiarity of the readers with the topics. We show that the three most important features are textual features, references and images.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Computer-supported cooperative work, Evaluation/methodology, Web-based interaction* ; K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*

General Terms

Experimentation, Human Factors, Reliability, Verification

1. INTRODUCTION

1.1 Wikipedia

Wikipedia is becoming an increasingly popular source of encyclopedic information. In August 2009, it was listed as the sixth most popular website[1], which is far ahead of sites for similar purposes (e.g. Encyclopaedia Britannica: rank 3,050). With the increasing use of Wikipedia, the need for a reliable assessment of the trustworthiness of the presented information is also rising. Since Wikipedia is the "free encyclopedia that anyone can edit" one can never be entirely sure about the true source of information. One well-known example is the case of Professor Ryan Jordan, well respected for his contributions to Wikipedia. As revealed in an interview with the New Yorker[2] he turned out to be a 24-year old community college drop-out.

1.2 Information quality

A concept closely related to information trust is information quality. Kelton et al.[14] describe trust as playing a key role as a mediating variable between information quality and information usage. Hence, trust can be seen as an assessment of the information quality on which the decision whether to use the information is based.

Despite controversies like the example in section 1.1, the quality of the articles on Wikipedia has proven to be high. In 2005, Nature compared 42 articles on scientific topics to the matching articles in the Encyclopaedia Britannica[11]. Wikipedia articles contained four errors on average whereas Encyclopaedia Britannica contained three. However, this study resulted in much criticism, especially from the publishers of Britannica who claimed the study was "fatally flawed"[3] by its defects in the methodology of the comparison. Nature responded that these flaws favored neither Wikipedia nor Britannica. It would be interesting to see an updated comparison since the articles in both encyclopedias have undergone another four years of development since Nature's original evaluation. The Wikipedia community also has grown enormously in size (17413 user accounts in January 2005 versus 539973 user accounts in October 2009[3]), indicating that more people have been working on the articles.

Conversely, it has been shown that the work needed to maintain the high quality is rapidly increasing[15]. In the early days most edits were adding information to the articles. However, nowadays increasing effort is going into reverting vandalism (the intentional destruction of an article) or so called 'edit wars' (constantly reverting the changes of another user because of a dispute about a topic). Vandalism is often easy to detect when, for example, an entire article is replaced by a single - often idiotic - phrase. It becomes harder when just a few numbers, dates or facts are altered. 'Edit wars' can also be hard to detect when lay people read an article, since such 'edit wars' are mostly the result of two

[1] www.alexa.com/topsites

[2] www.newyorker.com/archive/2006/07/31/060731fa_fact

[3] stats.wikimedia.org/EN/TablesWikipediaEN.htm

WICOW'10, April 27, 2010, Raleigh, North Carolina, USA.

or more editors with different perspectives on a topic. Such disputes should, according to Wikipedia's guidelines, be resolved on the corresponding discussion page, but this is not always done.

1.3 Information sources

Due to the open character of Wikipedia, the true source of most of the information remains unknown. However, attempts have been made to give more insight into the nature of the authors. WikiScanner[12] maps the IP-addresses of contributors to organizations across the whole world. This has revealed the particular interest of some organizations (such as government institutes) in the information provided on Wikipedia.

Anthony et al.[2] took another approach. They defined two types of users with distinct characteristics of their contributions. "Good Samaritans" are registered users who regularly add and correct information. The quality of their contributions increases over time as they become more experienced. The other group is referred to as "Zealots". They are unregistered users (only identifiable by their IP-addresses) who add or correct information only occasionally. The quality of their first contribution is often high (for example correcting a typo or an obvious flaw) but the quality decreases over time.

The unknown source of information is a problem when trying to apply conventional information trust measures[6] to the articles. A large portion of such measures is based on evaluating characteristics of the information source. Other aspects on which online trust is built include the design of graphics, structure, content and social cues[22]. Fogg et al. [10] have determined that the three most important features of websites on which credibility is based are design, structure and focus. However, most of the aspects found in these studies are standardized throughout Wikipedia and hardly differ among articles. Therefore, they are of no value in assessing trustworthiness.

1.4 Assessing trustworthiness

Other ways of coping with the trustworthiness of information in articles thus need to be taken. There are several views on this subject. The first to discuss is that of the Wikipedia Editorial Team (WPET). As stated earlier, information trust and quality are very closely related. The WPET assesses the quality of articles on Wikipedia manually; ranking each in one of the seven available classes (see Table 1). These classes are well-defined with clear rules about which features should appear in articles in a certain class. Note that only summaries of the descriptions are listed in Table 1; more detailed descriptions are available on their special page[4]. Important aspects in this evaluation include style, structure, illustrations, factual accuracy, stability, neutrality, length, and comprehensiveness. As of April 2009, the Wikipedia Editorial Team have assessed over 80% of all the articles. These assessments are very labor-intensive which creates problems in view of the ever-increasing number of articles and edits. Note that the classifications are normally not presented on the same page as the article itself (except for the highest achievable classification and occasionally the lowest). Instead, the class in which the article is categorized is stated on the discussion page. Most of the assessed articles are listed as stub (68.36%) or start (24.77%)

[4]en.wikipedia.org/wiki/Wikipedia:1.0/A

Status	Description
FA	The article has attained Featured article status.
A	The article is well organized and essentially complete, having been reviewed by impartial reviewers from a WikiProject or elsewhere. Good article status is not a requirement for A-Class.
GA	The article has attained Good article status.
B	The article is mostly complete and without major issues, but requires some further work to reach Good Article standards. B-Class articles should meet the six B-Class criteria.
C	The article is substantial, but is still missing important content or contains a lot of irrelevant material. The article should have some references to reliable sources, but may still have significant issues or require substantial cleanup.
Start	An article that is developing, but which is quite incomplete and, most notably, lacks adequate reliable sources.
Stub	A very basic description of the topic.

Table 1: Wikipedia Editorial Team Assessment

articles. Only a few C (1.42%), B (2.91%), GA (0.39%), A (0.02%) and FA (0.14%) articles are available, but these tend to be on the more important topics on the WPET importance scale[4].

A second way to look at trustworthiness is through the eyes of regular contributors. Potential edits to articles are often discussed before (and afterwards) on the discussion pages. Stvilia et al.[19] have looked at the elements of articles which are often discussed on these pages. The main aspects are: accessibility, accuracy, authority, completeness, complexity, consistency, informativeness, relevance, verifiability, and volatility. These aspects are much the same as those of the Wikipedia Editorial Team.

An important observation is that both the regular contributors and the Wikipedia Editorial Team are only small subsets of all Wikipedia users. Just a small percentage of visitors to the website contribute to the articles. Over 50 percent of edits are made by less than 1 percent of users[20].

Since most users are visitors who do not add or change any content, it is interesting to see which elements of articles they use to assess their trustworthiness. Similarities with the aspects used by the Wikipedia Edit Team and Stvilia et al.[19] are expected, but distinct features might also be discovered.

Once the elements used by lay readers have become clear, they can be used as a basis for a heuristic measure of trustworthiness. This measure can be presented to visitors to help them in their assessment of trustworthiness. We hypothesize that the advantage of using heuristics of the same elements as lay readers would use is that the measure will be easier to understand. This is in contrast to algorithmic measures that already have been developed by various researchers (for example [1], [5], [23], [7], [13]), but which are much harder to understand by the user.

1.5 Experiment

We performed an experiment in which lay readers assessed the trustworthiness of Wikipedia articles. During the task, the participants were instructed to think aloud[9]. The aspects on which they based their assessment were extracted from the think aloud protocol and classified.

Our study is comparable to that by Fogg et al.[10]. As noted above, they asked visitors to many websites about those features on which they based their assessment of credibility. The main difference is that in our research, all articles come from the same website which means that several features described in Fogg's study do not apply. Examples of such features are company motive, identity of the site sponsor, and customer service.

The articles used in this experiment were selected such that the familiarity of the participants with the topics varied. In most studies, experts judge the articles (e.g. [4],[11]). Normally, when people visit Wikipedia they will not be experts on a topic since they want to learn more about it[17]. We expect the participants will rate articles on topics they are already familiar with as more trustworthy[4]. Variation in features on which the assessment is based might also be found. It is for example likely the participants will be able to use their own knowledge more with familiar topics.

The second manipulation is in the quality of the articles. The quality of Wikipedia articles ranges from stubs (just short descriptions) to featured articles (comprehensive, well-written articles). To replicate the varying quality of articles on Wikipedia, articles of various classes as defined by the Wikipedia Editorial Team are used in the experiment. We expect that distinct features are used for articles of both high and low quality, as seen in the assessments of the Wikipedia Editorial Team.

1.6 Hypotheses

The main hypothesis concerns the features used by lay readers in the assessment of trustworthiness of Wikipedia articles. We expect that although there will be an overlap with the editors' features and the WPET, distinct aspects of trustworthiness will be found in this research. This is partly due to a different mental model on the importance of the various available cues. Also, the task performed by the Wikipedia Editorial Team differs from the task performed in this experiment. Namely, the WPET assesses quality whereas in this experiment trustworthiness is rated. This renders certain features useless, such as copyright issues of images.

HYPOTHESIS 1. *The features used by lay Wikipedia readers overlap with those of the Wikipedia Editorial Team, but different features will also be used.*

A second expectation is that distinct features are used when assessing good versus poor quality articles. Also, some features might be more salient than others when looking specifically at positive or negative comments by participants. This leads to the following two hypotheses:

HYPOTHESIS 2. *The features used by lay Wikipedia readers differ for articles of good and poor quality.*

HYPOTHESIS 3. *The features used by lay Wikipedia readers differ for positive and negative comments on an article.*

We expect that people will be able to use their own knowledge more effectively when assessing articles on familiar topics, enabling them to use content features (such as the correctness of the text).

We also hypothesize that when the familiarity of the topic of an article is high, there will be a strong positive bias on the trustworthiness, as seen in earlier research[4]. Readers will likely be confirmed on some of their own knowledge, which gives them confidence in the trustworthiness of all information. This leads to the following hypotheses:

HYPOTHESIS 4. *The features used by lay Wikipedia readers differ for articles on familiar and unfamiliar topics.*

HYPOTHESIS 5. *The trustworthiness ratings of the participants are higher for articles on familiar topics than on unfamiliar topics.*

As stated earlier, people will be able to verify the actual content more in articles on familiar topics. We expect that this verification will mean that the assessments take longer.

HYPOTHESIS 6. *People take more time to assess articles on familiar topics than on unfamiliar topics.*

Next, the method of obtaining the features used by lay Wikipedia readers in an experiment is presented. After this the results from the experiment are given, followed by a discussion and suggestions for future research.

2. METHOD

2.1 Participants

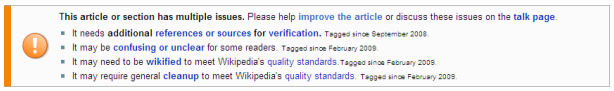
Fifteen university students took part in the experiment. Three of them had to be excluded from the analysis, two due to their poor performance on the think aloud task and one due to a technical problem regarding the audio recording. The average age was 23.4 years ($SD = 6.3$). Seven participants were Dutch and five were German. They received course credits for their participation. All participants spoke proficiently Dutch. None of them had problems expressing their thoughts in Dutch, so this language was chosen for the think aloud method[8]. Their experience with Wikipedia ranged from 3-8 years with an average of 5. All participants were able to explain the basics of Wikipedia in their own words; none of the participants had experience in editing articles on Wikipedia.

2.2 Task

We refer to the task performed in this experiment as the 'Wikipedia Screening Task'. In this task a Wikipedia article was opened in a web browser on a 17" computer screen. The appearance of the article suggested that they were looking at Wikipedia itself. In fact, the articles were off-line versions which were slightly manipulated to remove any cues of trustworthiness or quality. Examples of these cues are small bronze stars as seen in featured articles or infoboxes as shown in Figure 1. Infoboxes signal flaws in the quality in the articles. The participants were not aware of the modifications made to the article. Also, they were not allowed to visit any other web pages.

The articles were taken from the English Wikipedia. Although no native English speakers took part in the experiment, the participants did not report major language barriers.

Figure 1: Infobox containing information on quality



The participants were asked to rate the trustworthiness of each article. The way to form a view of the trustworthiness was not specified, so they were free to choose and develop their own methods. The assignment was specifically to rate the perceived trustworthiness and not to assess other factors such as relevance or entertainment value.

During the task, the participants were instructed to think aloud according to standard think aloud instructions [9]. Using this method, everything that passed through their minds had to be verbalized. The experimenter who was present during the session did not interrupt the participant while thinking aloud. No time limit was set.

2.3 Design

The design of the experiment was 2 (familiarity) × 2 (article quality). Familiarity and article quality were within subjects factors. The order of familiarity was alternated, starting with a familiar topic. The order of article quality was randomized.

2.4 Independent variables

2.4.1 *Familiarity*

Familiarity was manipulated to look for differences between the assessment of articles on familiar and unfamiliar topics. The topics were selected after a short interview with each participant, held a few days before the experiment. In this interview, participants were asked about their areas of interest, hobbies and expertise and areas they were not interested in or knew little about. The articles were then selected by hand for each participant. No article appeared more than once during the experiment which means that 120 unique articles were used, 10 for each participant.

2.4.2 *Article quality*

Article quality was manipulated following the classifications used by the Wikipedia Editorial Team. Six of the seven available classes (see Table 1) were used since only few A-class articles appear on Wikipedia. For each participant, within both the familiar and unfamiliar condition, no quality level appeared more than once.

The six classes were divided into good and poor quality. Featured articles, Good articles and B-class articles were considered good quality whereas C-class articles, start articles and stub articles were considered poor articles.

2.5 Dependent variables

2.5.1 *Protocol analysis*

The audio of entire experiment was recorded. Afterwards, the protocol was typed out as plain text. Based on this text, phrases containing comments on the trustworthiness were selected. The comments were then categorized and classified as being positive, negative or neutral. The coding scheme was created during the analyses of the first few participants, with extra categories and features added when required.

The experiment was conducted by two experimenters. Six protocols were analyzed by each experimenter, five of them being from experiments he led himself and one of an experiment led by the other. Based on the resulting two double-coded protocols, the inter-rater reliability was calculated. Cohen's Kappa was .79, which marks a substantial agreement[16].

The protocols of the twelve participants were combined by averaging the percentages of the various features. This was done to prevent participants with a large number of comments to have more influence on the results than participants with fewer comments.

The protocols of the various conditions were compared by using Chi Square-tests on the main feature categories.

2.5.2 *Trustworthiness ratings*

After viewing each article, the participant was asked to rate the trustworthiness on a 7-point Likert scale.

2.5.3 *Motivations for the trustworthiness ratings*

Besides giving trustworthiness ratings, the participants were asked to write down the aspects on which their ratings were based. These aspects could either contribute positively or negatively to the rating. Their motivations were categorized and compared to the results of the protocol analysis. Differences between the results of these two techniques might suggest that the participants were partially unaware of the features they used in their trustworthiness assessments.

2.5.4 *Familiarity ratings*

To check the manipulation of the familiarity with the topics, the participants were asked to rate their familiarity on a 7-point Likert scale.

2.5.5 *Trial duration*

The duration of each trial was measured in seconds to check for differences between familiar and unfamiliar articles.

2.6 Procedure

The experiment began with a questionnaire on the participants' familiarity with Wikipedia and various demographic features. After completing the questionnaire, brief instructions were given on the Wikipedia Screening Task and the think aloud method.

The participants practiced these tasks with two practice articles before the experiment began. The topics were 'Flat Earth'[5] and 'Ethnography'[6], and had Wikipedia Editorial Team classifications of 'good article' and 'start article', respectively. After the practice session, the performance of the participants on both the Wikipedia Screening Task and the think aloud task was considered sufficient to start the experiment.

The participants were presented with ten Wikipedia articles. Half the articles were on a topic they were familiar with, whereas the other half were on an unfamiliar topic. After they indicated that they had finished assessing the current article, a short questionnaire was presented. In this questionnaire they were asked to rate the trustworthiness

[5] en.wikipedia.org/wiki/Flat_Earth
[6] en.wikipedia.org/wiki/Ethnography

(with a motivation on positive and negative aspects) and their familiarity with the topic of the article.

The experiment ended with a few control questions about the manipulations. The duration of the experiment was roughly 90 minutes.

3. RESULTS

3.1 Manipulation checks

The familiarity of the participants was varied by manipulating the topics of the articles. The questionnaires after each article confirmed the manipulation. On a -3 to 3 familiarity scale the participants rated the familiar articles ($M = 1.66, SD = 0.63$) higher than the unfamiliar articles ($M = -2.35, SD = 0.59$); $t(11) = 16.08, p = 0.00$.

The manipulation of article quality is confirmed by higher trustworthiness ratings in the good quality condition ($M = 1.76, SD = 0.52$) than in the poor quality condition ($M = 0.52, SD = 1.05$); $t(11) = 4.08, p = 0.00$.

3.2 Extraction of comments

A total number of 1147 comments were extracted from the think aloud protocols. From the questionnaires, 456 comments were extracted.

Table 2 shows the results of both the protocol analyses and the questionnaires. A significant difference in the distribution of the comments over the categories was found between the two methods ($\chi^2(10) = 20.28, p < 0.05$). This is most likely due to the large difference in numbers of comments on textual features (26.33% versus 48.68%). Our further analysis is based on the results of the think aloud protocols. The advantage of this method is that it gives more information about the cognitive processes involved in this task compared to questionnaires[9].

3.3 Overlap with Wikipedia Editorial Team

The features most often noted by the participants in the experiment were textual features, references and pictures. A statistical difference was found between the distribution of comments over categories in this experiment and a chance-based distribution ($\chi^2(10) = 32.83, p = 0.00$). An overview of all features found is shown in Table 2.

Earlier we determined that the most important features used by the Wikipedia Editorial Team to judge quality include style, structure, images, references, stability, neutrality, length, and comprehensiveness.

Most of these features were also regularly stated by the participants in the experiment. Most frequent were references and images. Style, structure and comprehensiveness were also used in the assessment of trustworthiness.

However, the WPET does also use unique features which were not observed in the experiment. Stability is a factor which was not noted by the participants but was important for the Wikipedia Editorial Team. Neutrality was also hardly noted by the participants. One feature that was not used by the WPET but is stated by the participants is the use of internal links.

This confirms hypothesis 1: There is an overlap on the features used by lay readers and the WPET, but different features are also found. No statistical evaluation of this comparison has been performed since the WPET did not actually take part in this experiment. The reason that these

	Prot.	**Quest.**
Appearance	**4.97%**	**5.04%**
General	2.27%	0.00%
Structure	2.70%	5.04%
Tab. of Cont.	**4.62%**	**0.88%**
General	3.66%	0.44%
Length	0.52%	0.22%
Structure	0.35%	0.00%
Contents	0.09%	0.22%
Introduction	**5.06%**	**1.54%**
General	2.18%	0.66%
Length	0.70%	0.66%
Clarity	1.05%	0.22%
Contents	1.13%	0.00%
History S.	**3.57%**	**1.54%**
General	2.35%	1.32%
Length	0.44%	0.00%
Clarity	0.26%	0.00%
Contents	0.52%	0.22%
Infoboxes	**1.39%**	**0.00%**
General	1.05%	0.00%
Relevance	0.09%	0.00%
Clarity	0.00%	0.00%
Overview	0.26%	0.00%
Lists	**2.70%**	**1.54%**
General	2.35%	1.32%
Relevance	0.09%	0.00%
Clarity	0.09%	0.22%
Overview	0.17%	0.00%
Pictures	**12.55%**	**9.65%**
General	5.06%	1.97%
Relevance	2.44%	2.19%
Captions	0.17%	0.22%
Quality	3.31%	2.85%
Quantity	1.57%	2.41%
References	**26.07%**	**24.78%**
General	8.98%	2.19%
Relevance	1.05%	0.00%
Quality	6.45%	6.80%
Quantity	9.59%	15.79%
Int. links	**5.84%**	**6.36%**
General	3.75%	2.19%
Relevance	0.61%	0.22%
Quality	0.00%	0.66%
Quantity	1.48%	3.29%
Text	**26.33%**	**48.68%**
General	0.09%	0.00%
Scope	1.31%	2.85%
Writing style	1.48%	4.39%
Neutrality	1.22%	3.73%
Clarity	2.62%	6.58%
Comprehen.	6.36%	18.64%
Correctness	9.94%	7.24%
Length	3.31%	5.26%
Other	**6.89%**	**0.00%**

Table 2: Coding scheme with all features mentioned in the think aloud protocol (Prot.) and questionnaires (Quest.)

two groups are compared qualitatively is that we take experts on Wikipedia as a baseline for expected features to be used by lay readers.

3.4 Good and poor quality

No statistical difference was found between the distribution of comments over the categories between comments on good and poor quality articles ($\chi^2(10) = 3.62, p = 0.97$). Therefore hypothesis 2 cannot be confirmed.

Post hoc inspection of the data shows that in the good quality condition more general comments were made on pictures, whereas the poor quality condition contained more comments on most features within the textual feature category.

3.5 Positive and negative comments

Hypothesis 3 has to be rejected when comparing the distribution of comments over the categories ($\chi^2(10) = 13.05, p = 0.14$).

However, post hoc inspection does reveal some tendencies. The feature category 'text' was noted more for negative comments than for positive comments. Almost every feature in this category was noted more, but the biggest difference was in comprehensiveness and length. However, correctness was noted more often as a positive comment.

Comments on references were also more often negative than positive. This was mainly caused by comments on the number of references.

Comments on appearance were predominantly positive. This seems to be mainly caused by positive remarks on the general appearance of an article ("this article looks good to me"). No negative remarks on the general appearance were recorded.

3.6 Familiar and unfamiliar topics

Hypothesis 4 can also not be confirmed since no statistical difference in distributions was found ($\chi^2(10) = 4.22, p = 0.95$).

Post hoc inspection shows that some features seem to have been used specifically in articles on familiar or unfamiliar topics. The best example of this is the feature correctness. Since you have to know something about the topic in order to be able to judge the correctness of information, this feature was virtually only observed for articles on familiar topics. Other textual features, where prior knowledge of the topic is not needed, were used more when confronted with unfamiliar topics. Examples of these are clarity, comprehensiveness and length. References and pictures were also stated more for unfamiliar topics.

3.7 Trustworthiness ratings

No significant difference was found in the trustworthiness ratings of articles on familiar topics ($M = 1.28, SD = 0.75$) and unfamiliar topics ($M = 1.00, SD = 0.72$); $t(11) = 1.43, p = 0.18$. Therefore, we reject hypothesis 5.

3.8 Trial durations

The duration in seconds of the trials with articles on familiar topics ($M = 264, SD = 75$) was not significantly longer than with unfamiliar topics ($M = 250, SD = 79$); $t(11) = 1.37, p = 0.20$. Hypothesis 6 has to be rejected based on this result.

4. DISCUSSION

4.1 Trustworthiness ratings

In this research we tried to find out on which features trustworthiness assessments of lay Wikipedia readers are based. We found out that the most important categories are text, references and pictures. Manipulation of article quality or familiarity with the topic do not seem to influence the assessments of trustworthiness. Also, no differences between positive and negative comments were found in terms of used features. Only post hoc inspection of the data shows some tendencies towards features being more important in certain conditions.

An explanation for the lack of difference in comments on good and poor quality articles is the fact that even poor articles are of relatively good quality. This means that articles of lesser quality often look virtually the same as good quality articles in the sense that they share the same features (e.g. references, images). Having the same features, it is likely that comments on them are comparable. However, this is not confirmed by the trustworthiness ratings of good and poor quality articles, which are higher for good quality articles.

We also found no difference between the ratings of articles on familiar and unfamiliar topics. Familiarity does not seem to influence trustworthiness in this task and setting. This result is seemingly in contrast with the findings of Chesney[4], who found that experts have more trust in information than non-experts. However, note that the difference that he found was only significant at the 10% level. Next to this, Chesney used actual experts in the fields of the topics of the articles, whereas the participants in this experiment were mostly only familiar with the topic and not experts per se.

4.2 Features Used

The differences that can post hoc be observed within the three largest categories are discussed here.

Table 2 shows that the most recorded feature category was 'text'. This was partly because of the broad scope of the features in this category. However, some of the features in this category were noted quite often. Two very salient features were comprehensiveness and correctness. There seems to be a trade-off for these two features between the familiar and unfamiliar condition. Correctness was seen more often in the familiar condition whereas comprehensiveness was more salient in the unfamiliar condition.

The explanation for this difference is quite clear for correctness: since the participant will not be familiar with the topic in the unfamiliar condition, they will not be able to judge whether the given information is correct. In the familiar condition, the knowledge of the participants is often confirmed by the information in the article. This is also supported by the fact that the majority of the comments on correctness are positive.

The difference in the occurence of the comprehensiveness feature in the familiar and unfamiliar condition was slightly smaller than for the correctness feature. However, it is seen more in the unfamiliar condition. We hypothesize that this difference is caused by the fact that the participants do not know much about the topic in the unfamiliar condition. They are severly hampered when trying to understand the given information when it is incomplete. This is confirmed

by the fact that the majority of remarks on this feature are negative. In the familiar condition, this difference is much less salient.

A third feature that is quite often recorded in the 'text' category is length. However, given earlier findings on this subject by Palfrey and Gasser[18], we would have expected this feature to be seen more. One of their participants noted that if a piece of text is longer than a mobile text message, someone obviously put in the effort of writing it and it could not possibly be wrong. Length is seen slightly more in the unfamiliar condition, where participants have to assess the trustworthiness on a more superficial level due to a knowledge deficit.

Textual features are mentioned much more while assessing articles of poor quality. This difference is caused by an increase in comments on comprehensiveness and length. The features found are very similar to those used by the Wikipedia Editorial Team to indicate articles of poor quality. Note that this observation is not reflected in the questionnaires afterwards.

The number of occasions on which the feature category 'references' is noted is almost as high as text. Besides general remarks on this feature, the number of references is noted the most. The quality of the references is also often noted. We define the quality of references by their type, for instance, books, papers or websites. It is remarkable that this feature is noted more often in the unfamiliar condition since it seems easier to comment on the quality of references when being familiar with the topic.

We suggest that the type of participants in this experiment, namely academic students, are the cause of the exceptionally large number of remarks on references. When the experiment would be repeated with a different demographical group, this feature might be much less salient.

The third feature category that is noted very often is pictures. Next to the general remarks, relevance, quality, and quantity are noted.

Relevance is noted slighty more in the familiar condition, possibly due to the same effect as correctness in the 'text' category: the participants have an expectation based on their own knowledge and what they read in the text, which is then confirmed by a relevant picture.

The reason why quality is noted more often in the unfamiliar condition is less clear. It might be caused by the observation that mostly superficial features are used in this condition. In this case, these include remarks on the resolution, colors, or general esthetics, which are all categorized under quality.

The feature category 'pictures' is seen more in the assessment of good articles than poor articles. Poor articles often lack relevant pictures of good quality whereas good articles are mostly well-illustrated. The fact that more comments on pictures are seen for good quality articles may indicate that the appearance of pictures is noticed and valued, but the omission of them is less salient.

4.3 Differences with the WPET

We see that a lot of the features mentioned by the participants are also used by the Wikipedia Editorial Team. It is however important to stress that the task performed by the WPET is not the same as performed in this experiment. The WPET rates quality based on strictly defined guidelines, whereas in this experiment the perceived trustworthiness was rated without specifying the method to use for the assessment.

Because of this difference, some features assessed by the Wikipedia Editorial Team are not relevant or salient for lay readers. An example is stability. In this experiment, the participants did not have the chance to see how stable the contents of an article were. It is however highly questionable whether they would check this in a normal everyday setting.

The way features are evaluated also differs. For example both the participants in this experiment and the Wikipedia Editorial Team are looking at pictures. However, the participants are mainly looking at their visual appearance, whereas the Editorial Team will also assess other aspects such as copyright issues.

4.4 Trial Duration

We expected the participants to need more time to assess articles on familiar topics than on unfamiliar topics. However, the difference between these two conditions was not found to be significant. This is caused by very large individual differences of the participants resulting in a high standard deviation.

We hypothesize that two contrasting influences have led to this result. On the one hand, participants that are assessing articles on familiar topics are able to verify the correctness of information in the article, which we expect to take longer. On the other hand, when articles on unfamiliar topics are presented, it might take the participants longer to understand and comprehend the information in the article, also leading to a longer trial duration.

5. FUTURE RESEARCH

The task performed by the participants in this experiment was to assess the trustworthiness of Wikipedia articles. In this research the task of assessing trustworthiness was made very explicit. Although we think that one should always be careful with information from Wikipedia, in practice trustworthiness assessments are very minimal and implicit [21]. During this experiment, participants became more aware of the risk of untrustworthy information then they might normally be.

The results of the experiment are likely to be heavily dependent on the demographics of the participants. Since all of them were academic students, we expect an academic bias. This can for instance be seen by the impact of 'references' in the features used. We propose to repeat this experiment using a group with other demographical features. In our opinion it is important that this group has a general purpose to use Wikipedia (e.g. education). An example of such a group would be high school children.

Using the knowledge we now have on the features used by lay Wikipedia readers, research can be performed on the characteristics of these features. Experiments can be carried out in which particular features are manipulated according to theory-based hypotheses, for instance by evaluating the influence of the 'comprehensiveness' feature on perceived trustworthiness.

Another direction of research would be the development of heuristic decision support systems which help users to judge trustworthiness. We hypothesize that support systems which take the features users would utilize themselves into account will be more helpful than systems that use a complex algorithm not accessible to the user.

6. ACKNOWLEDGMENTS

We would like to thank Koen Remmerswaal for his efforts as a co-experimenter in this research. We are also grateful to Matthijs Noordzij and Maarten Hoeppermans for their helpful comments on preliminary versions.

7. REFERENCES

[1] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. Technical Report UCSC-SOE-08-07, School of Engineering, University of California, Santa Cruz, CA, USA, May 2008.

[2] D. Anthony, S. W. Smith, and T. Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of wikipedia. November 2005.

[3] E. Britannica. Fatally flawed: Refuting the recent study on encyclopedic accuracy by the journal nature. March 2006.

[4] T. Chesney. An empirical examination of wikipedia's credibility. *First Monday*, 11(11), November 2006.

[5] T. Cross. Puppy smoothies: Improving the reliability of open, collaborative wikis. *First Monday*, 11(9), September 2006.

[6] M. Domino and C. R. Cano. Inherited trust of information: A preliminary scale development. *SSRN*, -, August 2007.

[7] P. Dondio, S. Barrett, S. Weber, and J. Seigneur. Extracting trust from domain analysis: A case study on the wikipedia project. pages 362–373. 2006.

[8] K. Elekes. T̃ please, keep talkingT̃: The 'think-aloud' method in second language reading research. *NovELTy*, 7(3), October 2000.

[9] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. The MIT Press, 1984.

[10] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *DUX '03: Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15, New York, NY, USA, 2003. ACM.

[11] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[12] J. Giles. Wikipedia 2.0, with added trust. *The New Scientist*, 195(2622):28–29, September 2007.

[13] M. Hu, E. P. Lim, A. Sun, H. W. Lauw, and B. Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252, New York, NY, USA, 2007. ACM.

[14] K. Kelton, K. R. Fleischmann, and W. A. Wallace. Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59(3):363–374, 2008.

[15] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: Conflict and coordination in wikipedia. In *CHI*, 2007.

[16] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

[17] S. Lim. How and why do college students use wikipedia? *Journal of the American Society for Information Science and Technology*, 60(11):2189–2202, 2009.

[18] J. Palfrey and U. Gasser. *Born Digital: Understanding the First Generation of Digital Natives*. Basic Books.

[19] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. Information quality discussions in wikipedia. Technical Report ISRN UIUCLIS–2005/2+CSCW, University of Illinois, 2005.

[20] D. Tapscott and A. D. Williams. *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio Hardcover, April 2008.

[21] A. Walraven, S. Brandgruwel, and H. Boshuizen. How students evaluate information and sources when searching the world wide web for information. *Computers & Education*, 52(1):234–246, January 2009.

[22] Y. D. Wang and H. H. Emurian. An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior*, 21:105–125, 2005.

[23] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and Mcguinness. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, October 2006.