University of Twente

Bachelor thesis

Trust of high school students in Wikipedia

Author:
Andreas Bremer
s0116297

Supervisors:
Teun Lucassen
Matthijs Noordzij

July 1, 2010

Abstract

Wikipedia is an open encyclopedia: Nearly everyone can edit most of its content. Having been a key element for Wikipedia's striking success, the open character poses a challenge to its users: How can they form an opinion about whether to trust the information? Lucassen and Schraagen (2010) have made an approach to define the criteria that are used by academic students to evaluate the trustworthiness of Wikipedia articles. This is a follow-up study to further qualify the findings of Lucassen and Schraagen with another population, namely high school students. In an experimental setting, Familiarity and Quality of the articles are manipulated. We show that high school students employ different features than academic students to assess the trustworthiness of Wikipedia articles. The most frequently used features were textual features and pictures. It is concluded that when evaluating trustworthiness of Wikipedia articles, high school students mostly rely on comparison with their own knowledge.

Contents

1. Introduction

Wikipedia is a free online encyclopedia with growing importance for information retrieval. It is an open-content system using the MediaWiki software application[1], which makes it possible to edit most of it's content without restrictions, even anonymously. Only some articles are restricted from open editing and can only be edited by registered users.

In general, this principle means a great potential for knowledge distribution. But it can be both a blessing and a curse. At most times, it is not possible to retrace the origins or correctness of statements in Wikipedia articles. But how do humans judge the reliability of such statements; which criteria are used?

1.1. Related Work

There has been a lot of research on the perception of quality in Human-Computer-Interaction (Rieh & Danielson, 2007). The fundamental efforts of Fogg and Tseng (1999) and Rieh and Danielson (2007) have made approaches to define the concept of computer credibility. According to Fogg and Tseng, credibility is a *perceived quality* and made up of multiple dimensions. The authors emphasize the difference between the concepts of *trust* and *credibility*: *credibility* relates to the believability of information, while *trust* indicates the reliability of information, thus the willingness to depend on an information and it's correctness. They concluded that *credibility* consists of two major factors: *trustworthiness* and *expertise*. Rieh and Danielson (2007, p. 9) in turn found that, although labeling differs across studies, credibility is a notion with various underlying relevance criteria such as *expected quality*, *source quality*, *authority*, and *reliability*.

Various approaches to study the perceived quality in relation to Wikipedia have been made. Chesney (2006) conducted an experiment in which experts had to evaluate the credibility of Wikipedia articles. He found that experts judged articles from their field of expertise as more credible than those dealing with other issues. Chesney hypothesized that this may stem from a cynical attitude towards the Internet, but there was no difference found between the groups in terms of their cynicism. Additionally, there was no correlation found between the respondent's cynicism and perceived credibility. Having dismissed this as an explanation, Chesney concluded that experts found Wikipedia articles more credible than non-experts and that therefore the accuracy of Wikipedia's information is high.

Kittur, Suh, and Chi (2008) investigated whether and how the perceived reliability of Wikipedia can be changed. To do so, they conducted an experiment. Kittur et al.

[1] `http://nl.wikipedia.org/wiki/MediaWiki`

designed visualizations describing the editing history of selected articles and manipulated various metrics (percentage of words contributed by anonymous users, stability of the content, last edit made by an anonymous or established user and past editing activity) to develop two versions: high-trust and low-trust. These visualizations were then added to the articles. The quality of the articles was also varied by two levels, which were chosen out of the hierarchy defined by the Wikipedia Editorial Assessment Team[2] for evaluating Wikipedia articles. *Good article* (GA) class articles served as high quality articles while B-class articles (mostly complete and without major issues, but require some further work to reach *Good Article* standards) were used for the low quality condition. The results were compared with a baseline condition (the same experimental setup but without visualization of the editing history). An interaction effect was found, indicating that participants rated the quality of GA-class articles significantly higher when provided with a visualization than without. B-class articles were also rated lower when combined with visualization than without.

Pirolli, Wollny, and Suh (2009) did a comparable study but their visualizations of the history of Wikipedia articles were generated using a tool named WikiDashboard[3]. It actually works in real-time and on the "real" Wikipedia. Their results strengthend the findings of Kittur et al.: users' perception of credibility increased resp. decreased significantly more than in the baseline condition (without visualization). According to Pirolli et al. (2009), this effect could result from offering asymmetrical information (i.e., providing more information in one condition than in others). This effect was already studied decades ago within economics (Akerlof, 1970, as cited in Pirolli et al., 2009). It predicts that when more information about a product's quality is given, the quality will be perceived as higher.

Adler and De Alfaro (2007) used the information present in article versions to compute a content-driven reputation for Wikipedia authors. Their reputation system for Wikipedia had prescriptive and descriptive value (i.e., it also allowed for direct inferences about the quality of freshly entered text). This was achieved through developing algorithms for computing content-driven reputation, which measured the statistical correlation between the author's reputation at the time an edit was done and the subsequent lifespan of the edit. Employing this method, Adler et al. (2008) created a system named WikiTrust[4], which visualized the trustworthiness of a Wikipedia article directly. Their idea to evaluate a trust labeling was that high trust should be associated with stability. The system

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

[3] http://wikidashboard.parc.com/

[4] http://wikitrust.soe.ucsc.edu/

of Adler et al. (2008) computed the trust of a word within a Wikipedia article based on the article's revision history and the reputation of the contributing authors (Adler & De Alfaro, 2007). The trust of each word was then indicated by coloring the background of the word: white for fully trusted words; increasingly intense gradations of orange for progressively less trusted text. In the meantime, the authors have further developed WikiTrust: An online version is available as plug-in for Mozilla Firefox.[5]

According to Metzger (2007), there are myriads of possible factors that can contribute to the evaluation of credibility. In empirical studies, however, only few are found. Metzger (2007) compared the findings of Rieh (2002), Fogg et al. (2003) and Eysenbach and Kohler (2002): similar in all studies, credibility evaluations were influenced by the information itself (i.e., it's content, presentation, structure) and the source (i.e., it's reputation and type). But the weight of the different factors differs between the various studies. In studies with academic populations, source characteristics seem to be more important than, for example, the presentational features (Rieh, 2002). Instead, in the study of Fogg et al. (2003), *design* was the most named factor.

Agosto (2002, as cited in Rieh & Danielson, 2007) found that female high-school students (from the ninth and tenth grade) seem to rely more on design than on source quality when searching the internet for information. Specifically, they made strong positive responses to both the color and the design of graphics and multimedia. According to Agosto, perceived quality of information content proved to be a primary evaluation criterion, but her participants tended simply to equate information quality with information quantity.

Recently, Lucassen and Schraagen (2010) studied the evaluation of trust of academic students in Wikipedia articles. They did an experiment in which the trustworthiness of Wikipedia articles had to be assessed. Participants had to verbalize their thoughts (Think Aloud method (Ericsson & Simon, 1993)). Assessments were subsequently analyzed to see which features were used to evaluate trustworthiness. Quality of and familiarity with the articles were manipulated: the former to create a more realistic, representative situation, the latter to investigate whether there are differences in the assessment of familiar and unfamiliar topics. Lucassen and Schraagen found that the three most important features are textual features (25.59% of the comments made by the participants), references (26.93%) and images (13.13%). They concluded that the high proportion of references may stem from an academic bias, since all participants were students.

[5] https://addons.mozilla.org/en-US/firefox/addon/11087/

1.2. Experiment

The purpose of the present study was to investigate whether and how the distribution of assessment features changes if the demographical features were varied. To achieve this, we conducted a follow-up study to that of Lucassen and Schraagen (2010). The experimental setup was replicated, but our test population consisted of high school students with an average age of roughly 14.

The results of this study can further qualify the study of Lucassen and Schraagen (2010) in terms of allowing to further specify the influence of the population on the evaluation of trust in Wikipedia. Furthermore, they could possibly be of practical value when creating interactive environments for younger individuals.

1.3. Hypotheses

The main hypothesis concerned the difference between the two test populations and the distribution of the features used for assessing trustworthiness. We expected the distributional pattern in this study to be significantly different from the one found by Lucassen and Schraagen (2010).

Hypothesis 1. *The features used by high school students to assess the trustworthiness of Wikipedia are significantly different from those used by students.*

The following hypotheses are based on those stated by Lucassen and Schraagen (2010). It was expected that different features are used for the assessment of familiar and unfamiliar topics. For example, it is likely that people use their own knowledge to evaluate the correctness of articles. Furthermore, we expected that participants would rate articles within their field of familiarity as more trustworthy than unfamiliar ones. This would be comparable to the findings of Chesney (2006) but in contrast with those of Lucassen and Schraagen, who could not find evidence for hypotheses 2 and 3.

Hypothesis 2. *The features used to assess the trustworthiness of familiar articles differ significantly from those used at non-familiar articles.*

Hypothesis 3. *The ratings of the trustworthiness of familiar articles are significantly higher than ratings made upon non-familiar articles.*

We also measured the time needed to evaluate articles. The comparison with the own knowledge takes time, so Lucassen and Schraagen (2010) expected that more time is needed to rate the familiar articles, but no significant evidence was found to prove this. In spite of these findings we tested this for best comparability between the two studies, which led to hypothesis 4:

Hypothesis 4. *More time is needed to evaluate familiar articles than non-familiar ones.*

Hypotheses 5 and 6 also couldn't be proved by the results of Lucassen and Schraagen (2010), but were tested again for best comparability between the two studies. We presumed that for the assessment of articles with good quality, other features will be used more than for articles with bad quality. Additionally, we expected a different distribution of positive and negative comments over categories. For example, articles of bad quality usually had fewer pictures than articles of good quality, or even did not contain a single picture. It was therefore thought to be likely that pictures would not be mentioned when not present and further, that they would be mentioned less in a negative context. This led to hypotheses 5 and 6:

Hypothesis 5. *Different features are used for the assessment of articles with good quality than for articles with poor quality.*

Hypothesis 6. *Different features are used for negative and positive comments on an article.*

In the next section, the method of the experiment will be presented, followed by the results and a discussion.

2. Method

2.1. Participants

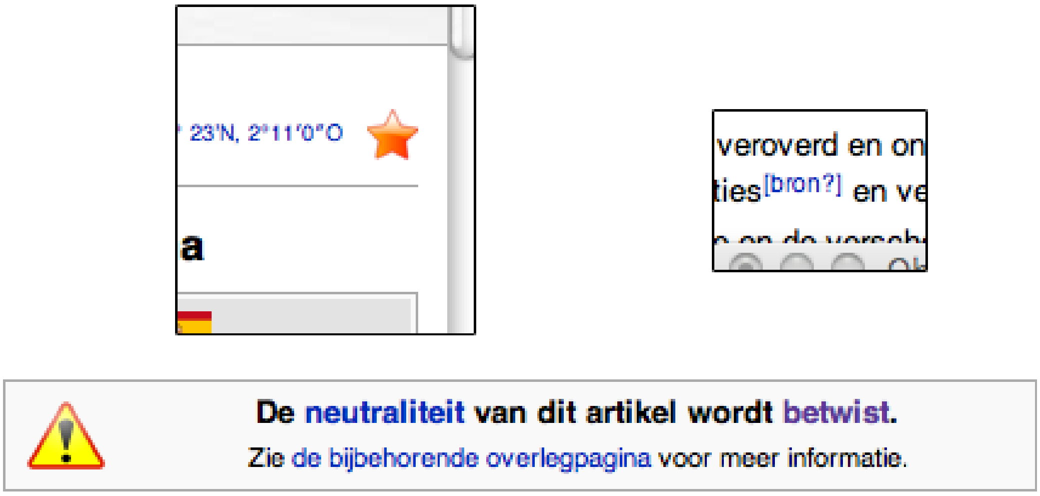
13 high school students (8 female, 5 male) with a mean age of 14,3 ($SD = 0,63$) took part in the experiment. All 13 were recruited from the VWO-3 niveau of the public school CSG Dingstede in Meppel, the Netherlands. They were all native dutch speakers. As compensation for their participation, they received 6 Euros. Their experience with Wikipedia ranged from 2 to 5 years ($M = 3.85, SD = 0.9$).

When asked to explain Wikipedia in their own words, most participants stated that it was an online encyclopedia, where information, definitions and pictures could be found.

The open character of Wikipedia was mentioned by only three participants. Only one had experience in editing articles.

2.2. Apparatus

The articles were presented on 15,4" laptops. Audio information was recorded during the whole experiment, using the software Audacity and an external USB-microphone. For each of the 13 participants, 10 individually selected Wikipedia-articles have been prepared. Preparation involved storage on the local hard drives and removing any cues to article quality (little stars indicating very good articles, warning boxes informing about possible flaws in neutrality, categorical hints; Fig. 1).



De neutraliteit van dit artikel wordt betwist.
Zie de bijbehorende overlegpagina voor meer informatie.

Figure 1: Examples of cues to article quality in the Dutch Wikipedia. These were systematically removed from our stimuli.

There was only one major difference between this study and the one of Lucassen and Schraagen (2010): due to the young age it was not certain if the english skills of the high school students were sufficient for such a sophisticated task. Consequently, we used the Dutch version of Wikipedia instead of the English version.

As the experiments were held in a public school in various rooms, the setup of the desks and chairs was changed if needed, so that the experimenter could sit next to the participant during the practice trials. Furthermore, a sign was attached to the door informing about the experiment and asking for silence. To assure equal procedures among both experimenters, a script (see section A.1) was created which contained all steps of the experiment in the right order. Furthermore, every participant received printed instructions (see section A.3). The three Questionnaires (Previous, During and Afterwards; depicted in section A.4) and the instruction sheet were taken from Lucassen and Schraagen (2010) and adapted to fit the requirements of the present study.

2.3. Task

Participants had to rate the trustworthiness of each article. They were told that they don't have to care about the relevance of an article or the entertainment value.

They were instructed to use the Think Aloud method (Ericsson & Simon, 1993; Van Someren, Barnard, Sandberg, et al., 1994). This method involves the verbalization of everything that passes through the mind. It requires a certain amount of practice, so there were two practice trials in the beginning. The assessment of the perceived trustworthiness was twofold: First, the total trustworthiness of an article was measured after each article by a questionnaire. Second, after the experiments, the Think Aloud protocols were analyzed to determine which features were used by the participants to assess the trustworthiness.

During the trials, only the experimenter was present. No time limit was set.

2.4. Design

The design of the experiment was 2 (familiarity: Familiar/Unfamiliar) x 2 (quality: good quality/poor quality) with familiarity and quality as within-subjects factors. The order of familiarity and quality was randomized independently over participants.

2.5. Manipulation

A few weeks before the experiment started, participants were contacted to ask if they would like to participate in the experiment. When they agreed, an appointment was made. Furthermore, they were asked to name five hobbies and/or fields of interest. Then, during the article preparation phase, 5 articles per participant were selected to match these topics (familiar articles) while the other five were taken from other, different topics (unfamiliar articles).

Articles were also manipulated with respect to their quality. Unfortunately, in the Dutch Wikipedia there is no such extensive categorization of article quality as the WikiProject article quality grading scheme in the English Wikipedia maintained by the Wikipedia Editorial Team. There is only a category with especially good articles (called the "Etalage") and a list of articles with different kinds of problems (e.g., missing neutral point of view). The quality of the articles had thus to be judged by the experimenters themselves. The WikiProject article quality grading scheme served as orientation for this assessment.

2.6. Procedure

After agreeing to participate in the present study, participants received an informed consent form (see section A.2), which they and their parents had to read and sign. When arriving, participants were greeted and had to fill in the first questionnaire. This contained items about some demographic features and various items about their knowledge and attitude towards Wikipedia.

When participants had filled in the first questionnaire, they received the instruction sheet and were instructed to read it carefully. Then the practice trials began. During these, participants were instructed and guided with respect to the use of the Think Aloud method. All participants used the same practice articles 'Barcelona'[6] and 'Titanic'[7].

In the actual experiment, each participants had to process ten articles: five of them were on familiar topics, whereas the other half were on unfamiliar topics. When participants were ready with an article, they had to indicate this on their own. After each article, they were asked to fill in a short questionnaire, in which they had to rate the trustworthiness of the article, give a motivation for this rating and rate their familiarity with the topic.

2.7. Analysis

2.7.1. Manipulation Checks

After each article, participants had to rate it with respect to trustworthiness and familiarity on a 7-point Likert scale. The use of the t-test by Lucassen and Schraagen (2010) to check for the effect of their manipulations in their study was criticized because 7-point Likert scales do not depict parametrical data, which this test actually requires. Therefore, manipulations were checked with the Wilcoxon signed-rank test (Moore & McCabe, 2001). This was done separately for both independent variables (familiarity and quality).

2.7.2. Protocol Analysis

The think-aloud protocols were typed into a plain text file. This text was split up into single statements and sorted into an Excel-spreadsheet. Each of the statements was then inspected, categorized and classified as positive, neutral or negative remark.

[6] http://nl.wikipedia.org/wiki/Barcelona_(Spanje)
[7] http://nl.wikipedia.org/wiki/Titanic_(schip)

The categories and coding scheme were adapted from Lucassen and Schraagen (2010). Protocol counts from all 13 protocols were merged 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 results were weighted and analyzed via Microsoft Excel, SPSS and GNU R using the χ^2-Test for Homogeneity (Huizingh, 2004; Moore & McCabe, 2001) and Fisher's exact test (Agresti, 1992).

2.7.3. Interrater Reliability

The experiment was conducted by two experimenters. Each one coded and rated his own 6 resp. 7 protocols. Additionally, two protocols were double coded and compared by each other. An interrater reliability analysis using Cohen's Kappa (κ) with 95% confidence interval was performed to determine consistency among raters (Sheskin, 2004, p. 543-547). It calculates agreement beyond chance. κ has a maximum of 1 when agreement is perfect, but a value of 0 indicates no agreement better than chance, and negative values show worse than chance agreement. According to Altman (1991, p. 404) the κ values are to be interpreted as in Table 1.

Value of κ	Strength of agreement
< 0.20	Poor
0.21 - 0.40	Fair
0.61 - 0.80	Good
0.81 - 1.00	Very Good

Table 1: Kappa Values

2.7.4. Additional Measurements

Besides the trustworthiness and familiarity ratings after each article, participants were asked to name aspects of the motivations on which their trustworthiness-ratings (positive and negative) were based. These were also coded and compared using the same method as in section 2.7.2 on page 11.

Additionally, trial duration for each article was measured to compute averages (familiar, non-familiar and total duration) and check them for differences.

3. Results

The interrater reliability for the raters was found to be $\kappa = 0.767$ (p < 0.001), 95% CI (0.638, 0.896). A statistical difference was found between the distribution of the comments across the 11 main categories and a chance-based distribution ($\chi^2(10) = 107.066, p < 0.0005$). The two most used features were textual features (75.83%) and pictures (11.27%).

3.1. Manipulation checks

Subsequent to each article, participants were asked to indicate their familiarity with the corresponding article topic on a 7-point Likert scale ranging from -3 to 3. High values indicated high familiarity. Familiar articles were rated significantly higher ($M = 1.06, SD = 1.076$) than non-familiar articles ($M = -1.36, SD = 0.875$); ($Z = -3.113, p < 0.001$).

The trustworthiness of good articles was also measured with a 7-point Likert scale ranging from -3 to 3, while high values indicated high trustworthiness. The trustworthiness ratings however did not confirm the manipulation with respect to article quality: participants rated articles of good quality with a mean of $M = 1.682$ ($SD = 0.703$) not significantly higher than articles of poor quality ($M = 0.717, SD = 1.852$); ($Z = -1.365, p = 0.086$). Averaged, only 7 of the 13 participants rated the trustworthiness of articles of poor quality lower than articles of good quality ($N = 7, M = 9.29$, sum of ranks = 65). This may result from the higher standard deviation in the ratings for articles of poor quality. To reassess the manipulation's effectivity, the quality of the articles was blind-rated again by colleagues, which yielded a kappa of $\kappa = 0.886$ (p < 0.001).

3.2. Extraction of comments

A total of 745 comments were extracted from the Think Aloud protocols, 229 from the questionnaires. No significant differences were found while comparing the distribution of comments across both methods ($\chi^2(8) = 3.893, p = 0.867$). There are only marginal differences between the percentages (Tab. 2 on page 14). While performing this test, it became clear that the Expected values for some categories were too low ($E < 1$). Therefore, the categories Infoboxes, Table of Contents, Internal Links and First Alinea were excluded from the analysis. The test was once more performed, which again did not yield any significant differences ($\chi^2(4) = 0.846, p = 0.932$). 60% of the cells had an expected value below 5, which marks a problem for the applicability of the χ^2-Test: it poses a violation of the thumb rule, that for $r \times c$-Tables greater than 2×2, less than 20% of the cells should have an expected value below 5 (Moore & McCabe, 2001, p.

507). To maintain comparability between the present study and the one of Lucassen and Schraagen (2010) we decided to stick to this method. The further analysis was based on the Think Aloud protocols.

Category	Quest.	Prot.
General Appearance	2,57%	3.00%
Table of Contents	1.15%	1.16%
First alinea	0%	1.34%
History section	2.39%	2.92%
Infoboxes	0%	0.36%
Lists/Tables	1.89%	1.45%
Pictures	9.36%	11.27%
References	0%	0%
Internal Links	0.27%	1.65%
Textual features	81.25%	75.83%
Other	1.12%	1.00%

Table 2: Comparison of Think Aloud protocols (Prot.) and questionnaires (Quest.)

3.3. Academic students and high school students

Comparing the distribution of comments over categories across high school students and academic students, there was highly significant evidence for a different distribution in both populations ($\chi^2(9) = 61.289, p < 0.0005$) (Fig. 2 on page 15). Infoboxes were excluded from the analysis because of the low expected values ($E < 1$). 70% of the cells had an expected value below 5. Therefore, Fisher's exact test was performed, resulting in a p-value of 8.414×10^{-14}, ($p < 0.001$). This confirmed hypothesis 1: Academic students and high school students make different use of features when assessing the trustworthiness of Wikipedia articles.

Taking a closer look at the percentages in Table 3 on page 16, it became clear that none of the high school students had actually mentioned the references, while 26.07% of the academic students' comments did. Contrary to this, 75.85% of scholar comments were related to textual features, as opposed to only 26.33% of the academic students' comments.

Post-hoc inspection of the data revealed that 41.69% of the high school students' comments on textual features stem from comments about the feature *correctness*. With a total count of 11 (1.15%), 13 (1.54%) and 11 (1.65%), *Table of Contents*, *First alinea* and *Internal links* were mentioned relatively less by the high school students than by the academic students (43 (4.62%), 55 (5.06%) and 62 (5.84%)).

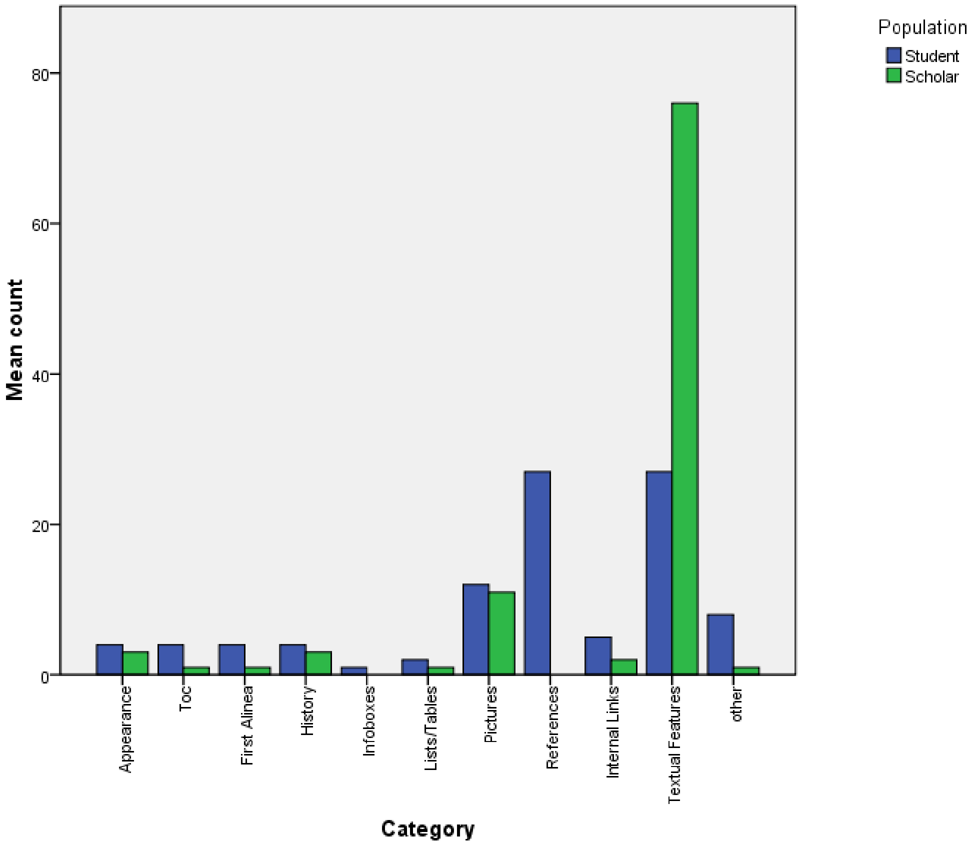


Figure 2: Distribution of comments over categories

	Acad.	**High s.**
Appearance	**4.97%**	**3.00**
General	2.27%	1.23%
Structure	2.70%	1.77%
Table of Cont.	**4.62%**	**1.15%**
General	3.66%	1.15%
Length	0.52%	0%
Structure	0.35%	0%
Contents	0.09%	0%
Introduction	**5.06%**	**1.54%**
General	2.18%	0.66%
Length	0.70%	0%
Clarity	1.05%	0.20%
Contents	1.13%	0.23%
History	**3.57%**	**2.92%**
General	2.35%	2.15%
Length	0.44%	0.59%
Clarity	0.26%	0.10%
Contents	0.52%	0.08%
Infoboxes	**1.39%**	**0.36%**
General	1.05%	0.23%
Relevance	0.09%	0%
Clarity	0%	0.13%
Overview	0.26%	0%
Lists/Tables	**2.70%**	**1.45%**
General	2.35%	1.19%
Relevance	0.09%	0%
Clarity	0.09%	0.13%
Overview	0.17%	0.13%

(continuation)	**Acad.**	**High s.**
Pictures	**12.55%**	**11.27%**
General	1.05%	7.60%
Relevance	2.44%	1.17%
Captions	0.17%	0.36%
Quality	3.31%	1.78%
Quantity	1.57%	0.36%
References	**26.07%**	**0%**
General	8.98%	0%
Relevance	1.05%	0%
Quality	6.45%	0%
Quantity	9.59%	0%
Internal links	**5.84%**	**1.65%**
General	3.75%	0.29%
Relevance	0.61%	0.13%
Quality	0%	0%
Quantity	1.48%	1.23%
Text	**26.33%**	**75.85%**
General	0.09%	9.49%
Scope	1.31%	0.96%
Writing style	1.48%	6.19%
Neutrality	1.22%	0.36%
Clarity	2.62%	4.25%
Comprehen.	6.36%	7.10%
Correctness	9.94%	41.69%
Length	3.31%	5.81%
Other	**6.89%**	**1.00%**

Table 3: Coding scheme with all features and corresponding categories mentioned by academic students (Acad.) and high school students (High s.)

3.4. Familiar and non-familiar topics

Comparing the distribution of comments over categories across familiar and non-familiar topics, no significant evidence was found for a different distribution in the two conditions ($\chi^2(9) = 11.877, p = 0.22$). The categories Infoboxes, Table of Contents, Internal Links and First Alinea had expected values below 1 and were excluded from the analysis. The test was performed once more, which again did not yield any significant differences ($\chi^2(6) = 8.833, p = 0.183$). 71.4% of the cells had an expected value below 5. Therefore, Hypothesis 2 had to be rejected: The features used for assessing the trustworthiness of articles covering familiar topics do not differ from the features used for the assessment of articles covering non-familiar topics.

Salient differences in the percentages (Tab. 4 on page 18) include *pictures* with a higher count in the non-familiar articles (13.32%) than in familiar ones (8.22%) and *textual features* with a higher count in familiar articles (78.94%) than in non-familiar ones (70.28%). Post-hoc inspection of the data showed that mostly in non-familiar topics the general presence of pictures was noted (and rated as positive or neutral), while the higher value of textual features is based on comments comparing the content positively with participants' own knowledge (i.e., there are many remarks about correctness of statements based on what participants already knew).

The trustworthiness ratings of familiar ($M = 1.412, SD = 1.095$) and non-familiar ($M = 1.031, SD = 0.957$) articles were analyzed using the Wilcoxon signed-rank test. 10 Participants rated the trustworthiness of familiar articles higher than the trustworthiness of non-familiar articles ($N = 10, M = 7.00$, sum of ranks = 70), which marks a significant finding ($Z = -1.717, p = 0.043$). This confirmed hypothesis 3: The ratings of the trustworthiness of familiar articles are significantly higher than ratings made upon non-familiar articles.

3.5. Trial durations

The mean duration of evaluating one article in seconds was $M = 165.31$ ($SD = 74.944$). There was no significant difference between the duration of familiar articles ($M = 173.08, SD = 89.981$) and non-familiar articles ($M = 175.46, SD = 70.423$); ($t(12) = -0.153, p = 0.441$). Therefore, hypothesis 4 (More time is needed to evaluate familiar articles than non-familiar ones) had to be rejected.

Category	Familiar	Non-fam.
General Appearance	2.13%	3.61%
Table of Contents	0.33%	1.76%
First alinea	1.10%	1.42%
History section	5.29%	0.37%
Infoboxes	0%	0.63%
Lists/Tables	0.76%	2.03%
Pictures	8.22%	13.32%
References	0%	0%
Internal Links	0.74%	2.65%
Textual features	78.94%	70.28%
Other	2.53%	3.94%

Table 4: Distribution of comments for familiar and non-familiar topics

3.6. Good and poor quality

When comparing the distribution of comments made on articles of good and poor quality, no significant evidence was found for a different distribution in the two conditions ($\chi^2(8) = 5.447, p = 0.709$). 77.8% of the cells had an expected value below 5. Therefore, hypothesis 5 had to be rejected: no difference in features used for the assessment of articles with good quality than for articles with poor quality could be found.

Taking a closer look at the percentages in Table 5 on page 19 reveal the only bigger difference: 12.84% of the comments made upon articles with good quality were related to pictures, while only 5.74% of the comments made upon articles of bad quality were related to pictures.

3.7. Positive and negative comments

There were no significant differences in the distribution of positive and negative comments ($\chi^2(8) = 10.042, p = 0.262$). Infoboxes, Table of contents, First alinea and Lists/Tables were excluded because the expected values were below 1. Subsequently, no differences were found ($\chi^2(4) = 4.658, p = 0.324$). 60% of the cells had an expected value below 5. Therefore, hypothesis 6 had to be rejected.

Post-hoc inspection of the data shows that no negative comments were made on Table of contents, First alinea and Infoboxes (Tab. 6 on page 19). Pictures were mentioned

Category	Good	Poor
General Appearance	2.94%	2.71%
Table of Contents	0.93%	1.62%
First alinea	0.94%	2.12%
History section	2.40%	3.27%
Infoboxes	0.36%	0.32%
Lists/Tables	0.67%	3.27%
Pictures	12.84%	5.74%
References	0%	0%
Internal Links	1.03%	2.65%
Textual features	74.02%	74.66%
Other	3.83%	3.63%

Table 5: Distribution of comments for good and poor article quality

more within a positive context (10.87%) than within a negative context (3.83%).

Category	Positive	Negative
General Appearance	3.57%	1.52%
Table of Contents	1.55%	0%
First alinea	1.65%	0%
History section	1.40%	2.60%
Infoboxes	0.59%	0%
Lists/Tables	1.42%	0.27%
Pictures	10.87%	3.83%
References	0%	0%
Internal Links	1.52%	1.60%
Textual features	72.08%	71.65%
Other	5.36%	10.84%

Table 6: Distribution of comments for positive and negative comments

4. Discussion

4.1. Distribution of comments

The present study was conducted to test the findings of Lucassen and Schraagen (2010) on another population, namely high school students. Therefore, the main objective was to assess the distribution of comments over categories and compare both populations (Hypothesis 1). The two most important features used by high school students to evaluate the trustworthiness of Wikipedia articles were textual features (75.83%) and pictures (11.27%). Lucassen and Schraagen found that the three most important features used by academic students were textual features (25,59%), references (26,93%) and pictures (13,13%). They concluded that the high proportion of references may stem from an academic bias. Since none of the high school students actually mentioned the references in any context, this conclusion is confirmed by this study. It is also in line with the findings of Rieh (2002), who found that in studies with academic populations source characteristics seem to be more important than presentational features.

The enormous proportion of the category textual features is based on the feature correctness. High school students apparently compared the content of the articles with their own knowledge. This can be confirmed when taking into account the proportions of positive and negative comments made about correctness: the number of positive comments about correctness was over twice as high as the number of negative comments. Post-hoc inspection of the data showed that textual features were also mentioned more in articles with familiar topics than in non-familiar ones, accompanied with a higher frequency of the feature correctness for familiar topics. The explanation for this seems logical: Because participants are familiar with the topics, they can judge the correctness of the information in familiar articles better by simply comparing it with their own knowledge.

Contrary, the features length and comprehensiveness were noted more often in a negative context. The explanation for this finding may be simply that high school students complained about shorter and incomprehensive articles. This would be in line with the conclusions of Agosto (2002) that high school students tend to equate information quality with information quantity.

Although not significant, there was a tendency that pictures were noted more in articles with good quality than on articles with bad quality. This may be based on the simple fact that articles of poor quality mostly contain less pictures than articles of good quality. Evidence for this assumption could be found when looking at the protocols: absence of pictures was only mentioned 3 times. Pictures were also mentioned more often in a positive context than in a negative one, which is in line with the findings of Agosto

(2002), whose participants (high school students about the same age as the participants in this study) made positive responses to both the color and the design of graphics and multimedia.

4.2. Manipulating and assessing trustworthiness

The possibility of manipulating the trustworthiness of Wikipedia has already been demonstrated (Kittur et al., 2008; Pirolli et al., 2009). However, the manipulation of trustworthiness purely through selection of articles by means of familiarity and quality did not produce measurable differences yet: neither in the study of Lucassen and Schraagen (2010) nor in the present study, at least in terms of statistical significant differences in the distribution of comments across categories. Post-hoc inspection of the data however revealed some tendencies in terms of the usage of particular features in certain conditions, as showed above.

The manipulation of trustworthiness in our study by means of varying article quality was not confirmed by the trustworthiness ratings of the participants, caused by 6 of the 13 participants who rated articles of low quality on average as more trustworthy than articles of good quality. This in turn may be a result of the fact that the Dutch version of Wikipedia was used—it contains less articles (604,972; 1-June-2010) than the english version (3,309,403; 1-June-2010). The pool of articles of good quality was therefore smaller, and compromises had to be made during the selection of articles, which could have led to this outcome. The blind reassessment of the article quality by colleagues though showed that differences in article quality were clear-cut, so the use of the Dutch Wikipedia appears not to be the reason for the found distribution of the trustworthiness ratings. Likewise, it might be the case that in high school students, the mental model of trustworthiness of Wikipedia is not yet as mature as in academic students. This would mean that high school students' understanding of trustworthiness consists mainly of comparison with already known facts and the presence of some neat pictures. Support for this conclusion can be found when taking into account the explanations of Wikipedia in the high school students' own words: Only three out of thirteen mentioned that Wikipedia can be edited by nearly anybody, which paraphrases Wikipedia's inherent, open character—a key element of the online encyclopedia.

Manipulation of familiarity instead led to the desired effects: significantly higher ratings on the familiarity scale. Furthermore, articles on familiar topics were rated as more trustworthier than articles on non-familiar ones (hypothesis 3). This outcome was in line with the results of Chesney (2006) but contrary to the outcome in the study of Lucassen and Schraagen (2010). This could be caused by the same reason implied from the assumption made above: since high school students seemingly evaluate Wikipedia's trustworthiness mainly by comparison with the own knowledge, it is not surprising that familiar articles were rated as more trustworthier.

4.3. Trial duration

No significant differences were found in the duration of evaluation of familiar articles compared to non-familiar articles. This may be due to the high standard deviation, which in turn is supposably induced by the large individual differences between the participants. When looking at these differences, one should keep in mind that the population consisted of adolescents, which possibly implies large individual developmental differences.

4.4. Further research

This study established that the assessment of Wikipedia's trustworthiness by high school students differs from the one made by academic students. However, some of the differences may be due to the age difference (about 9 years) between the two populations. Therefore, it would be of scientific value to conduct a comparable study with a population of about the same age as the academic students (or older, as developmental differences at this stage of life tend to diminish), but with a non-academic background, and compare the findings with the present study and the one of Lucassen and Schraagen (2010).

Another proposal would be to test this study with a population of native english speaking adolescents of about the same age as in the present study. Thereby one would be able to use the english version of Wikipedia and control for a factor, which formed a difference between the present study and the one of Lucassen and Schraagen (2010).

4.5. Conclusion

With the growing importance of Wikipedia as a source of information, it is also likely that more and more adolescents make use of it. This study showed that the assessment of Wikipedia's trustworthiness differs remarkably from that made by academic students. These findings imply that it is of great importance to teach critical thinking as early as possible or at least, make it clear to high school students that Wikipedia consists out of contributions of others with varying quality and that it must not be taken at face value.

References

Adler, B. T., Benterou, J., Chatterjee, K., De Alfaro, L., Pye, I., & Raman, V. (2008). Assigning trust to wikipedia content. In *Proceedings of the 2008 international symposium on wikis (porto, portugal).*

Adler, B. T., & De Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *Www '07: Proceedings of the 16th international conference on world wide web* (pp. 261–270). New York, NY, USA: ACM.

Agosto, D. E. (2002). A model of young people's decision-making in using the Web. *Library & Information Science Research*, *24*(4), 311–341.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, *7*(1), 131–153.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, *84*(3), 488–500.

Altman, D. G. (1991). *Practical statistics for medical research.* Chapman & Hall/CRC.

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. *First Monday*, *11*(11).

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Rev. ed.).* Cambridge, Ma: MIt Press.

Eysenbach, G., & Kohler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *British Medical Journal*, *324*(7337), 573.

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on designing for user experiences* (pp. 1–15).

Fogg, B. J., & 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* (p. 87).

Huizingh, E. (2004). *SPSS 12.0 voor Windows en data entry.* Den Haag, Academic Services.

Kittur, A., Suh, B., & Chi, E. H. (2008). Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. In *Proceedings of the acm 2008 conference on computer supported cooperative work* (pp. 477–480).

Lucassen, T., & Schraagen, J. M. (2010). Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on information* *credibility* (pp. 19–26).

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.

Moore, D. S., & McCabe, G. P. (2001). *Statistiek in de Praktijk* (3e herziene ed.). Den Haag, Academic Services.

Pirolli, P., Wollny, E., & Suh, B. (2009). So you know you're getting the best possible information: a tool that increases Wikipedia credibility. In *Proceedings of the 27th international conference on human factors in computing systems* (pp. 1505–1508).

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, *53*(2), 145–161.

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual review of information science and technology*, *41*, 307.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures.* CRC Pr I Llc.

Van Someren, M., Barnard, Y., Sandberg, J., et al. (1994). *The think aloud method: A practical guide to modelling cognitive processes.* Citeseer.

A. Materials

A.1. Script

1. Greeting
2. Explain what is going to happen in 1 sentence
3. Ask for informed consent form
4. Ask participants to deactivate their mobile phones
5. Questionnaire “Previous”
6. Hand out instruction sheet
   a. Check whether participants fully understood the instructions
   b. If not, repeat instructions
   c. Clarify that questions are allowed, but only during practice trials
   d. Clarify that there are no wrong answers
7. Start audio recording
8. Practice trial
   a. Clarify that participants have to indicate when they are ready with an article by themselves
   b. When handing out the questionnaire “During”, remind participants that this will be handed out after every article
9. After Practice articles:
   a. Recapitulate
   b. Give suggestions for better performance
10. Experiment:
    a. Talk as little as needed
    b. Offer glass of water
    c. Reduce interaction to a minimum
11. Stop audio recording
12. Hand out questionnaire “Afterwards”
13. Ask for account number
14. Debriefing

Naam onderzoeker: Andreas Bremer
E-mailadres: a.bremer@student.utwente.nl Telefoon (mobiel): 0049-xxx-xxxxxxxx

Verklaring van toestemming

Beste deelnemer,

U hebt toegezegd vrijwillig mee te doen aan een onderzoek. In dit document staat informatie over uw rechten en de procedure van het experiment. Lees de volgende paragrafen met veel aandacht.

1) Doel van het onderzoek

Het doel van dit onderzoek is meer inzicht te verkrijgen in hoe Wikipedia, een online encyclopedie, op betrouwbaarheid beoordeeld wordt door scholieren.

2) Procedure tijdens het experiment

In dit experiment laten we u 10 Wikipedia-artikelen zien op een computer monitor. Deze moeten door u op betrouwbaarheid beoordeeld worden. Na elk artikel wordt u gevraagd om een kort vragenlijst (3 items) te beantwoorden.
Het experiment duurt ongeveer één uur. De onderzoeker zal gedurende het gehele experiment bij u zijn en beschikbaar zijn voor vragen en opmerkingen.

3) Risico's en neveneffecten

Het experiment is goedgekeurd door de ethische commissie van de Universiteit Twente en is veilig en pijnloos voor de proefpersonen. Door mee te doen aan dit experiment zal u geen risico's ondervinden.

4) Beëindiging van het experiment

U hebt het recht om op elk moment het experiment te beëindigen zonder opgaaf van reden. De deelname is volledig vrijwillig en zonder verplichtingen. Er zijn geen nadelen verbonden aan beëindiging van het experiment uwerzijds.
Tijdens het experiment hebt u de mogelijkheid om een pauze te nemen. Als u tijdens het experiment een pauze wilt of mogelijk het toilet wilt bezoeken is dat mogelijk op elk moment. Als u (op welk moment dan ook) zich niet op uw gemak voelt, informeer dan onmiddellijk de onderzoeker.

5) Privacy en discretie

De beperkingen en regels die van toepassing zijn op de geheimhouding van de data worden gerespecteerd. Persoonlijke kenmerken zullen nooit ofte nimmer doorgespeeld worden aan derden. De verzamelde data wordt geanonimiseerd en zal enkel in voorgenoemde vorm geanalyseerd en gepubliceerd worden.

6) Verklaring

Door hieronder uw handtekening te zetten verklaart u het eens te zijn met onderstaande verklaring.

"Ik, de ondertekende, verklaar hierbij dat de onderzoeker me geïnformeerd heeft over bovenstaande punten. Ik heb de verklaring van toestemming gelezen en begrepen. Ik ben het eens met alle bovenstaande punten. Ik geef bij dezen toestemming om de te verzamelen data in geanonimiseerde vorm te laten analyseren en publiceren voor wetenschappelijke doelen. Ik ben op de hoogte gesteld van mijn rechten als proefpersoon en mijn vrijwillige participatie in dit onderzoek."

…………………………………………………………………..………………
Plaats Datum Handtekening

…………………………………………
Als de proefpersoon een minderjarige is, is de handtekening van een ouder of wettelijke voogd vereist.

Uitleg van het experiment

Gedurende dit experiment krijg je een aantal Wikipedia artikelen te zien. Stel jezelf voor dat je deze artikelen nodig hebt om een opdracht te schrijven. Voordat je de artikelen hiervoor gebruikt, wil je natuurlijk weten, hoe betrouwbaar de artikelen zijn.

In dit experiment zal jij dus de **betrouwbaarheid** van de artikelen beoordelen. Hoe je dat precies doet, bepaal jij. Je hebt onbeperkt de tijd om de artikelen te bestuderen. Je krijgt geen inhoudelijke vragen over de artikelen. Alleen de betrouwbaarheid is in dit onderzoek van belang.

De experimentleider zal vertellen welk artikel op welk moment geopend en bekeken mag worden. Het is niet toegestaan om te klikken op de pagina of om naar een andere pagina te gaan. Je mag wel scrollen om het hele artikel te kunnen bekijken. Je mag zelf aangeven wanneer je klaar bent met een artikel. Na elk artikel krijg je een korte vragenlijst over jouw oordeel over de betrouwbaarheid ervan. In totaal krijg je tien artikelen te zien. Indien gewenst kan er nog een korte pauze worden ingelast.

Terwijl je de taak uitvoert, zeg je alles wat je denkt, leest of doet hardop. Tijdens het invullen van de vragenlijsten hoef je dit niet te doen. Praat tijdens het experiment zo min mogelijk met de experimentleider.

Het hele experiment wordt opgenomen. Het verkregen materiaal is alleen voor wetenschappelijke analyse bestemd. Achteraf zal het niet mogelijk zijn te bepalen, dat de opname van jouw is.

Het experiment zal ongeveer een uur duren. Je krijgt nu eerst de gelegenheid om even te van de taak en het hardop denken te oefenen. Je krijgt hiervoor een voorbeeldartikel te zien. Wat bij deze oefening uitkomt wordt niet verder bekeken.

PP#
……..

Questionnaire vooraf

Voordat we aan het experiment beginnen worden enkele vragen gesteld over jezelf, Wikipedia en vertrouwen.

Geboortedatum:	……………………
Geslacht:	M / V
Nationaliteit:	……………………

1. Hoe lang geleden heb je Wikipedia leren kennen?

……… jaar

2. Hoe vaak maak je gebruik van Wikipedia?

Iedere dag	Iedere week	Iedere maand	Ieder jaar
□	□	□	□

3. Indien je bij vraag 2 "Iedere dag" hebt gekozen: Hoeveel uren besteed je per dag aan het gebruik van Wikipedia?

Meer dan 4 uur	Meer dan 2 uur	Meer dan 1 uur	Minder dan 1 uur
□	□	□	□

4. Leg zo goed mogelijk in eigen woorden uit wat Wikipedia is en hoe het werkt.

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

…………………………………………………………………………………

5. Met welk doel zoek je doorgaans informatie op Wikipedia?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

6. Heb je zelf al eens informatie toegevoegd of veranderd op Wikipedia?

ja / nee

7. Welke versie van Wikipedia heeft jouw voorkeur?

a. De Nederlandse
b. De Engelse
c. De Duitse
d. Anders, namelijk: ……………………………

8. Heb je informatie van Wikipedia wel eens rechtstreeks gebruikt in een opdracht of een opstel?

ja / nee

Indien ja, geef een voorbeeld:………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

9. In hoeverre vind je informatie van Wikipedia normaal gesproken betrouwbaar?

Zeer onbetrouwbaar			Neutraal			Zeer betrouwbaar
□	□	□	□	□	□	□

PP# …….. Art# ……..

Questionnaire na artikel

Op deze vragenlijst laat je jouw oordeel over de betrouwbaarheid van het voorgaande artikel weten. Wees opnieuw zo eerlijk mogelijk.

1. Hoe betrouwbaar kwam dit artikel op jouw over?

Zeer onbetrouwbaar			Neutraal			Zeer betrouwbaar
□	□	□	□	□	□	□

2. Waarop is je oordeel gebaseerd?

Positief:…..…………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

Negatief:……………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

3. Hoeveel wist je van te voren al over dit onderwerp?

Zeer weinig			Neutraal			Zeer veel
□	□	□	□	□	□	□

PP#

……..

Questionnaire achteraf

Hartelijk dank voor je deelname aan dit experiment. Als laatste willen we je nog wat vragen stellen over je deelname aan dit experiment.

1. In hoeverre kwam de taak van het beoordelen van de betrouwbaarheid die je tijdens dit experiment hebt uitgevoerd overeen met de manier waarop je normaal gesproken informatie op Wikipedia zou behandelen?

Zeer anders			Neutraal			Zeer gelijk
□	□	□	□	□	□	□

2. Merkte je veel verschil in hoeveel je over het onderwerp van de verschillende artikelen wist?

Zeer weinig			Neutraal			Zeer veel
□	□	□	□	□	□	□

3. Merkte je veel verschil in betrouwbaarheid tussen de artikelen?

Zeer weinig			Neutraal			Zeer veel
□	□	□	□	□	□	□

4. Opmerkingen.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

1 Chi-square tests

Computation example for familiar vs. non-familiar topics: is there a significant difference in distributions of comments?

H$_0$: The same features are used to evaluate familiar and non-familiar articles.
H$_A$: Different features are used to evaluate familiar and non-familiar articles.

1. Percentages were averaged using the Excel-spreadsheets "percentages_..."
2. Collected all percentages for both conditions in one file spss-file (familiarity2.sav)
3. Computed Chi-square value:
   There are $k = 11$ categories with $j = 2$ conditions.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \chi^2_{(k-1)(m-1);1-a} \quad (1)$$

$$= 11.877 \quad (2)$$

   In SPSS:
   Data → Weight cases → Weight cases by aantal
   Analyze → Descriptives → Crosstabs
   → Row variable: Familiarity
   → Column variable: Categories

Because no comments relating to references were made at all, they were removed from the table before computing the test statistics. Because the expected values of ToC, First Alinea and Infoboxes were below 1, these were excluded from the analysis. The result $\chi^2(6) = 8.833, p = 0.183$ gives no evidence for the alternative hypothesis, so H$_0$ cannot be rejected.

2 Wilcoxon signed-rank test

Mean ratings per participant were computed and entered into a new file ("wilcoxon_....sav"). Then the test statistics were computed: Analyze → Non-parametric tests → 2 related samples.

3 Fisher's exact test

The computations were made in R as showed in the following listing. First, the matrix in table 1 is entered (lines 1,2). Then, Fisher's exact test was computed (line 4). The p-value is shown in line 9.

```
 1 > distribution <- matrix(c(5, 3, 4, 1, 5, 1, 4, 3, 1, 0, 3, 1, 13,
       11, 26, 0, 6, 2, 26, 76, 7, 1), nr=2,
 2 + dimnames=list(c("Students", "Scholars"), c("Appearance", "ToC",
     "First Alinea", "History", "Infoboxes", "Lists/Tables", "
     Pictures", "References", "Internal Links", "Textual Features",
     "other")))
 3
 4 > fisher.test(distribution, workspace=2e6)
 5
 6         Fisher's Exact Test for Count Data
 7
 8 data:  distribution
 9 p-value = 8.414e-14
10 alternative hypothesis: two.sided
```

	Appearance	ToC	First Alinea	History	Infoboxes	Lists/Tables
Students	5.00	4.00	5.00	4.00	1.00	3.00
Scholars	3.00	1.00	1.00	3.00	0.00	1.00

	Pictures	References	Internal Links	Textual Features	other
Students	13.00	26.00	6.00	26.00	7.00
Scholars	11.00	0.00	2.00	76.00	1.00

Table 1
Matrix "distribution", containing distribution of comments across students and scholars