Influence of the first impression on credibility evaluation of online information

Bachelor thesis
Faculty of Behavioral Sciences
Psychology
University of Twente, the Netherlands

Niklas von der Goltz
s0202282

August 2012

1st Supervisor: Teun Lucassen, MSc.
2nd Supervisor: Prof. Dr. Jan Maarten Schraagen

WIKIPEDIA
The Free Encyclopedia

Table of contents

Abstract

With the success story of the internet during recent years and more and more people gaining information from a wide diversity and variety of websites (an estimated 644 million in 2012 according to Netcraft), credibility evaluation of online information has become an area of interest. On the other hand, within the field of cognitive psychology, a lot of research was done on intuition and "the power of the first impression" to get to know which aspects play a role in decision making and judgment of information. This study combines these two fields of knowledge and examines the influence of the first impression on credibility evaluation of online information. Furthermore, the 3S-model of credibility evaluation developed by Lucassen and Schraagen (2011) was integrated in the study, as it explains how people form their judgments on the credibility of information.

The main research tool in this study was Wikipedia, because of its importance and its status as one of the most important online information sources on the World Wide Web. Therefore, participants evaluated Wikipedia articles of different quality and familiarity in two time conditions; one short time condition lasting five seconds, and one long time condition lasting 5 minutes.

In this study, it was found out that when forming a first impression about an online article, more surface features than semantic features are applied for peoples' trust. However, in this study, the previous experience with Wikipedia as a source of online information had no influence on the trust in the articles of the study. Furthermore, quality manipulation revealed that peoples' trust is higher in the long time condition than in the short time condition for high-quality articles, but not for low-quality articles. No effect of time and familiarity on trust was found. However, an effect of quality on trust was found.

1. Introduction

During the last couple of years, a lot of research has been done on trust in online information and on the features people find important for their trust judgment (Kelton, Fleischmann, & Wallace, 2008; Risto, 2010; Lucassen & Schraagen, 2011). Among many others, the role of semantic features (Lucassen, Muilwijk, Noordzij, & Schraagen, in press), which are features based on the content of the text, and surface features (Lucassen et al., in press; Fogg, Soohoo, Danielson, Marable, Stanford, & Tauber, 2003; Robins, Holmes, & Stansbury, 2010; Lucassen, & Schraagen, 2010), which are features including references, pictures, length, or appearance, was examined. Moreover, research was done about the influence of one's previous experiences with an online source (Lucassen et al., in press; Hilligoss & Rieh, 2008) and about situational factors, such as time pressure (Wathen & Burkell, 2002; Metzger, 2007), on trust judgment.

However, the main focus of this study is the concept "first impression" and a determination is made how strong its influence might be. The question is posed on which features people base their credibility evaluation of an online article when they form a first impression. Furthermore, it is clarified whether the previous experience with a source of online information influences peoples' credibility evaluation and whether there is a difference in precisely this credibility evaluation when the quality of the articles is manipulated. Consequently, we measure the factor "trust" through a procedure called credibility evaluation. In the following, the factor "first impression" and the previous research on the topic is summed up. After that, the factor "trust" and the online information source Wikipedia are described in greater detail; as in this study, Wikipedia was the research tool to answer our hypotheses.

1.1 The first impression

The central concept in this thesis is the "first impression". We are interested in which role the first impression of users plays in whether they find an article from a certain website, in our case Wikipedia, trustworthy. We examined on which features people base their decision while evaluating the article on credibility; and whether either surface features, such as an appealing layout, or semantic features, such as academic accuracy or neutrality, matter for their trust in the article. We further tested whether the previous experience of people has an influence on their trust. Finally, we examined whether the quality of the article plays a role in this context.

In order to approach the concept of the "first impression", we discuss findings from the literature to explain the influence of the first impression on decision making; thus in our case, to evaluate credibility. According to the "dual-processing theory", two systems are working simultaneously when someone is making a decision: the first is 'system one', which is also called the heuristic system, being an automatic and implicit process. 'System two', which is also called the analytic system, on the other hand, refers to the rational and critical thinking side (Evans & Curtis-Holmes, 2005). On the one hand, the heuristic system has to do with emotions, suggesting that a decision in a certain situation is in part depending on the emotions that someone felt during that situation. On the other hand, the analytic system has to do with cognition and includes reasoning, weighting of choices and considering the outcomes (Evans & Curtis-Holmes, 2005). According to the dual-processing theory, there is a competition of both processes of reasoning about which of them to use in the first place (Evans & Curtis-Holmes, 2005).

Other scientists agree with the idea that there are two types of cognitive processes, which play a role in judgment and decision making. Kahneman (2003) described 'system 1' (= intuition) and system '2' (= reasoning): whereas operations influenced by system 1 are effortless, fast, automatic and influenced by emotions and habits, operations influenced by system 2 are effortful and therefore slower, as well as well-thought-out.

Moreover, trust in online information could be influenced by whether a user tries to rationally and logically evaluate information on the website to check if it is trustworthy, providing that he or she hast the time to do so; or whether he or she relies on surface features like the length of the article or images, and previous experience with the source. In their "elaboration likelihood model of persuasion (ELM)", Caccioppo, Kao, Petty, and Rodriguez (1986) made an effort to explain how people form attitudes and eventually change them later. The authors distinguish between two routes to persuasion: when using the 'central route', one thinks logically and intensively about an idea, dedicating a great deal of thought. On the other hand, when using the 'peripheral route', and mostly when being relatively unmotivated or under time pressure, one sticks to preexisting opinions or surface qualities (Caccioppo et al., 1986). However, one can use surface features when taking the central route and doing a well-thought-out credibility evaluation as well, as for example through paying attention to the quality of the references (Lucassen & Schraagen, 2011).

When applying the research named above [the "dual-processing theory", the idea of "system 1 and system 2", and the "elaboration likelihood model of persuasion (ELM)"] to our study; similarities can be found with the concept of the "first impression". When people form a first impression while evaluating online information on credibility, they first automatically use their heuristic system (or 'system 1' and 'peripheral route' respectively) to gain an overview of how credible the information might be. Whether they use their analytic system (or 'system 2' and 'central route' respectively) afterwards or stick to their first impression depends on factors such as their motivation and willingness to evaluate credibility (Scholz-Crane, 1998; Metzger, 2007; Lim, 2009), whether they are under time pressure (Plessner & Czenna, 2010), their information skills (Lucassen et al., in press), language (Eysenbach and Kohler, 2002); or situational factors (Corritore, Kracher, & Wiedenbeck, 2003; Wathen and Burkell, 2002).

Intuition is a process of thinking, which is defined as "what we know without knowing how we learned it", and can "serve as a basis for judgments and decisions" (Betsch, 2010, p.4). Even if reflective processes are generally superior compared with more intuitive ones when it comes to judgment and decision making, there is evidence for the benefits of intuitive decisions as well (Plessner & Czenna, 2010). According to McMackin and Slovic (2000), whereas cognitive efforts such as reflection and analysis tend to be more valid and effective in analytic tasks; intuition has an advantage in tasks where one has to deliver a quick judgment, as it turned out that, in these tasks, "thinking about reasons for decision degraded the quality of judgment" (p.529). When people are under time pressure, they rely on intuitive decisions as well (Plessner & Czenna, 2010).

According to Smith and Kosslyn (2009), recognition is "process of matching organized sensory input to stored representations in memory" (p.70) and is of importance when people rate online information. In his "recognition-primed decision (RPD) model of rapid decision making", Klein (1993) investigated the influence of time pressure on decision making, a factor that is suitable to this study. The model describes how people make quick and effective decisions when they are confronted with complicated and demanding situations. According to the model, when making a decision, people scan a possible course of action, ask themselves whether its outcomes would make sense, take the circumstances of the situation into account and then select the first course of action that fits their evaluation. According to this model, there is a difference between experts and novices: experienced people are capable of making quicker decisions as their

recognition is much faster and they already experienced similar situations or problems in the past. Inexperienced people, on the other hand, take the first course of action that comes along or rely on 'trial and error' (Klein, 1993).

The factor "time" comes into account as well when doing research on the influence of the first impression on trust judgment. In a study by Chase and Simon (1973), participants were divided into three groups in terms of their chess skills (experts, amateurs and novices) and then had to look at different chess positions for five seconds. After that, they had to reproduce the chess positions. Here again, a difference between experts and novices was found. It turned out that the experts performed better on reproducing the chess positions than the amateurs and the novices. The authors concluded that the advantage of the experts has its origin in their familiarity with chess positions and that this fact allows them to encode and retrieve the positions more efficiently, in larger integrated 'chunks' of information (Chase & Simon, 1973). Even if this experiment was about the effect of expertise on memory and not about trust judgment in online information, the general principle is the same: people are capable of encoding, storing, and retrieving a lot of information when they can connect the information to knowledge stored in their long-term memory. Kahneman and Klein (2009) call this the "intuition of experts". Lucassen and Schraagen (2011) investigated the role of expertise as well when they conducted a study comparing people being experts and novices in automotive engineering. In this study, it became obvious that trust of domain experts was largely influenced by accuracy, whereas trust of novices generally stayed unvaried.

The theories discussed above differentiated between people that base their judgments and decisions on a reasoned evaluation and people basing their judgments and decisions on a more implicit and intuitive evaluation, as well as on previous experiences. This difference may also be applied to the topic of interest in the current study, namely credibility evaluation. The same distinction can be found in the 3S-model developed by Lucassen and Schraagen (2011) to comprehend how people form their judgments on the credibility of information. The model consists of three strategies to evaluate credibility and is applied to rate online information.

According to Lucassen et al. (in press), the three strategies are as follows: the first strategy is to make allowance for semantic features of the information, for example for its academic accuracy, freedom from errors, or its neutrality. To be able to use this strategy adequately, a certain level of domain expertise is demanded, as one has to check and review the information.

Consequently, in this strategy, the defining factor of credibility is factual accuracy (Lucassen et al., in press).

The second strategy of the 3S-model is to consider surface features of the information. Surface features answer the question of how the information is presented and examples are photos, graphics, pictures, layout, the length of the article, or the number of references. Previous studies have shown that some of the most essential indicators for credibility come under the heading of surface (Fogg, et al., 2003; Robins, et al., 2010; Lucassen & Schraagen, 2010). To take surface features into account, the user needs generic information skills, meaning that he or she knows how certain features contribute to the concept of credibility (Lucassen et al., in press).

The third strategy of the 3S-model takes the preconceived opinion and previous experiences with a particular source into account. This strategy is passive and not the actual information is evaluated, but only the source of (online) information (Lucassen et al., in press). In practice, this could mean, for example, that if someone finds Wikipedia in general highly reliable, he or she trusts the information from a Wikipedia article, because "it's Wikipedia after all".
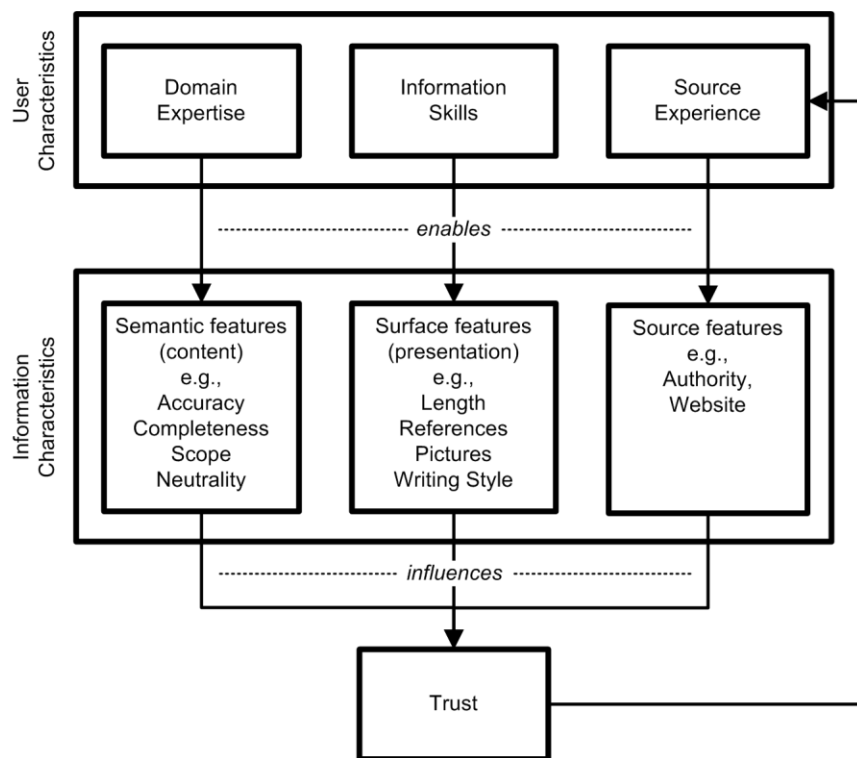

FIGURE 1. The proposed 3S-model of information trust as introduced by Lucassen and Schraagen (2011).

Consequently, the 3S-model underlines the other aspects that are of importance for this study, besides the influence of the first impression: semantic features, surface features, and the previous experience. In the following, we looked into the question in which case people tend to use semantic features, in which one surface features and which role their previous experience with a source of online information played.

In addition to that, there is more research on trust judgment of online information: for example, Eysenbach and Kohler (2002) state that the credibility evaluation is based on four elements: 1) whether the source is an "official" and well-received one, 2) whether the layout of the website is appealing and easy to use, 3) whether the author of the article supported his assertions with references from the academic literature, and finally, 4) whether the user was capable of understanding the language. According to Hilligoss and Rieh (2008), when people evaluate online information, they make three distinctive kinds of judgments: predictive judgments, evaluative judgments, and verification. During this process, they rely heavily on their accumulated knowledge and previous experiences (Hilligoss & Rieh, 2008). Summing up one can say that when they trusted Wikipedia before, they likely will have trust in any Wikipedia article appearing on the screen.

In this study, we used time manipulation to investigate whether an internet user's trust is changing if he or she has more time available. Consequently, the time variable had two levels: one, in which respondents had to look at a Wikipedia article for five seconds, the short time condition; and one, in which respondents had to look at a Wikipedia article for a period of five minutes, the long time condition. Respondents in the first setting presumably only had the time to gain a first impression and to check a small number of surface features; such as photos, graphics, pictures, layout, the length of the article, or the number of references. The respondents in the second setting; however, presumably had sufficient time to gain a substantial overview of the content of the article, to check some of the facts and to pay heed to things such as academic accuracy, spelling style, freedom from errors, or a highbrow diction. The study further examines which features (surface or semantic) people base their trust ratings on and whether their previous experiences with the online information source Wikipedia or the quality of the articles define their trust.

1.2 Trust

The question whether users trust digital and online information became more and more important during the last couple of years due to the expansion of the internet in our daily lives (Kelton et al., 2008; Naegele, 2008). Consequently, a definition of the complex construct "trust" had to be found, and this definition had to fit the online setting of the research as well. According to Rotter (1967), "trust" is an "[…] expectancy held by individuals or groups that the word, promise, verbal, or written statement of another can be relied on." (p. 651). For our example of the online encyclopedia Wikipedia, this definition fits perfectly, because, as people check the website for information, they have a certain level of trust (or mistrust) in the (in this case) written statement.

The appropriate kind of trust when it comes to online information, is interpersonal trust, thus the social connection between the trustor, the person who *trusts* (i.e, the user), and the trustee, the person or item who *is trusted* (i.e, Wikipedia) [Kelton et al., 2008]. The characteristics that define whether a trustee is to be trusted or not, are the trustee's perceived competence, goodwill, or future actions, among others (Kelton et al., 2008). While there were some objections on whether trust in a computer, or online information respectively, should be put on a level with trust in people from "real life", research showed that people are actually capable of trusting online information and that they even "relate socially to computer technology" (Kelton et al., 2008, p. 365). Even more so, according to Kini and Choobineh (1998), trust is the crucial aspect of whether people use a certain technology, in our case Wikipedia, again or whether they abandon it.

Besides trust, the term "credibility" is often used alternately in the literature, even if it is slightly different (Lucassen & Schraagen, 2011). Whereas credibility is a synonym for believability when it comes to a source or a message, trust means dependability (Fogg & Tseng, 1999). Consequently, sources of online information have a high credibility when users have the impression that the overall quality of their presented information is high. According to Hovland and Weiss (1952), the judgment of credibility depends on trustworthiness, whether someone is well-intentioned, and expertise, whether he or she is knowledgeable in a certain topic. Trust, on the other hand, is more active and involves the question whether someone is disposed to effectively depend on the credibility of information (Lucassen & Schraagen, 2011). However, this includes taking a certain risk that the information could be wrong, inaccurate, or incomplete.

Simply put: if a user finds the online information credible, his or her trust increases. In this study, we consequently measured trust through the process of credibility evaluation.

1.3 Wikipedia

Nowadays, more and more people are relying on Wikipedia as a tool for getting information about various topics. This fact increases the already huge importance of the website within the academic community and contemporary society as a whole (Tancer, 2007). Founded in 2001 by Jimmy Wales, an Internet entrepreneur, and Larry Sanger, a philosopher, Wikipedia has since then grown extensively to become the largest and most popular general reference work on the Internet (Miliard, 2008; Alexa, 2012). Moreover, it is the sixth most popular website worldwide, with an estimated 365 million readers and over 3.9 million articles on the English version alone (Woodson, 2007; Wikipedia, 2012).

It is no wonder that many students use Wikipedia to find background information and to acquire knowledge in a fast and comfortable manner (Lim, 2009; Jaschikk, 2007; Naegele, 2008; Wannemacher, 2009), but even experts and academic researchers are partly depending on the website (Fuchs, 2009). Features that make Wikipedia this attractive are its bandwidth of covered topics, the ease of use, and the fact that it is free, multilingual and accessible to everyone with an internet connection (Lim, 2009; Jaschikk, 2007). Consequently, the basic idea of Wikipedia is that every user is invited to join the community and to form and edit the knowledge database.

Therefore, and also because of the fact that there are no "official" or international formalities or rules for posting information online, the information can be inaccurate, completely wrong or misinterpreted by the user (Metzger, 2007; Völkel, Krötzsch, Vrandecic, Haler, & Studer, 2006; Walraven, 2008). However, the general information quality of Wikipedia is overall very high and the error rate is comparable with other encyclopedias, even if there is a certain "risk of low-quality information" (Lucassen & Schraagen, 2011, p. 1232). Furthermore, the website is assuredly not free from errors (Giles, 2005).

1.4 Hypotheses

Because of the fact that the goal of this study is to demonstrate the influence of the first impression on the credibility evaluation of online information, the concept "first impression" and related findings from the literature being of value for forming adequate hypotheses are described

in greater detail. According to Kahneman (2003), Plessner and Czenna (2010), McMackin and Slovic (2000), as well as Kahneman and Klein (2009), the concept of "intuition" is related to the first impression and its power on evaluation. Kahnemann (2003), for example, describes intuition as an important cognitive process for judgment and decision making and defined that decisions made with intuition as fast, automatic, influenced by emotions and habits, and effortless – all of which also applies to decisions relying on the first impression. Furthermore, the concept of "the first impression" is related to recognition (Smith & Kosslyn, 2009), a concept that is of relevance in our study when people rely on information stored in their memory for credibility evaluation.

To form our first hypothesis and to fit the "online information"-context of this study, we relate this research from the literature to the 3S-model, developed by Lucassen and Schraagen (2011), to get to know how people come to their credibility evaluation of online information. The authors underlined the importance of semantic features and surface features in this process. The fact that numerous studies either have shown that surface features are applied more for peoples' credibility evaluation than semantic features (Risto, 2010; Lucassen et al., in press) or underlined the general importance of surface features (Fogg et al., 2003; Robins et al., 2010; Lucassen & Schraagen, 2010) leads to the following first hypothesis:

Hypothesis 1: When forming a first impression about an online article, more surface features than semantic features are applied for peoples' trust.

In their 3S-model, Lucassen and Schraagen (2011) further address the importance of the previous experience with a source of online information for credibility evaluation. This fact, along with the work of Hilligoss and Rieh (2008), who found out that when people evaluate online information, they rely heavily on their previous experiences and accumulated knowledge with the source; and the work of Eysenbach and Kohler (2002), stating that credibility evaluation is partly based on whether the source is a well-received one, we formulate the following second hypothesis:

Hypothesis 2: The more positive peoples' previous experience with a source of online information, the higher is their trust.

According to Lucassen et al. (in press), the quality of online articles has a significant effect on trust, in the sense that people trust high-quality articles more than low-quality articles. Consequently, one can assume that people somehow recognize whether the information that they are confronted with is of high or low quality, at least to some extent. According with this finding and the time manipulation of this study, we wanted to investigate whether the factor time plays a role in this as well, or simply put: whether people score higher on trust if they have more time available for credibility information. People tend to rely on their intuition when they have little time for credibility evaluation of articles and, in contrast, tend to use their analytic system when they have sufficient time for credibility evaluation (Kahneman, 2003; Cacioppo et al., 1986; McMackin & Slovic, 2000; Plessner & Czenna, 2010). In line with this argumentation, we formulate the third hypothesis:

Hypothesis 3: Peoples' trust is higher in the long time condition than in the short time condition for high-quality articles, but not for low-quality articles.

2. Method

2.1 Participants

A total of 40 university students participated in the experiment. All of them were behavioral science students with the majority being either Psychology students or Communication Sciences students. The students enrolled online to take part in the study and received course credits in return. Other students were invited verbally or via e-mail. Of the 40 participants, 16 were male (40 %) and 24 were female (60 %) with the average age being 22.4 years (SD = 1.89). Moreover, 22 of them were Dutch (55 %) and 18 were German (45 %). The participants were familiar with reading articles in English, because of the fact that their courses of study were for a great part given in English.

2.2 Setting

Before the beginning of the actual experiment, the participants were asked to sign an informed consent. The participants received detailed instructions for the task verbally and in writing and had the possibility to ask questions. The Wikipedia articles were then displayed on a Computer screen.

To different time conditions were chosen: one short time condition and one long time condition. The two different time settings were chosen as follows: for the long time condition, the "5 min.-setting", it was important to make sure that the participants had no feeling of time pressure at all and that they had the possibility of looking long enough at the article. In a study by Risto (2010), the participants had unlimited time to read and to evaluate Wikipedia articles. It turned out that the mean viewing time per article was estimated at 4.46 minutes. Keeping this in mind, the time limit in our long time condition was set at five minutes. Additionally, at the end of the experiment, participants in the long time condition were asked whether they had the feeling of having had enough time.

For the short time condition, the "5 sec.-setting", it was essential that the participants were under time pressure while reading the Wikipedia article, but still being able to gain a first impression and to check a small amount of features. To assure this condition, a short pretest was conducted with N = 12, where participants were asked to have a look at a Wikipedia article and check some features. It turned out that in a "5 sec.-setting", people were under huge time pressure, but still had the chance to gain an overview and a first impression of the article. In an early stage of this study, we considered to use a "3 sec.-setting". However it turned out, that within three seconds, participants of the pretest had no possibility at all to check surface features and even the time for scrolling downwards was too limited. For this reason and to avoid the problem of long page loading times preventing participants from checking surface features, the "3 sec.-setting"- idea was dismissed.

2.3 Task

In this task, participants were solicited for performing the Wikipedia Screening Task, as described by Lucassen and Schraagen (2010). During the task, participants had to read Wikipedia articles and evaluate them on trust afterwards. Given the huge relevance of Wikipedia these days (Tancer, 2007; Miliard, 2008; Alexa, 2012; Woodson, 2007), the study concentrated on reading and rating articles from Wikipedia. The English version of the website was chosen, because this version, compared to other languages, contains a quality assurance by the Wikipedia Editorial Team (2012). Then, the two groups, the short time condition and the long time condition, were formed randomly.

Before the beginning of the actual task, participants anonymously had to give biographical data about their age, nationality, and gender. Then, their previous experience about the credibility of Wikipedia was examined briefly. Doing this, they first had to rate their overall trust in Wikipedia with a scale ranging from 1 ("not credible") to 7 ("credible"). To measure these ratings, a seven-point Likert scale was used, because of its overall high reliability and validity (Preston & Colman, 2000; DeVellis, 2012). Further, they then had to answer three additional questions to examine their opinion about Wikipedia including three "yes/no"-questions ("*Did you ever edit something on Wikipedia yourself?", "Did you ever use information from Wikipedia for an assignment at university?", and "Has it ever been the case that your trust in Wikipedia proved to be unjustified*?"). Moreover they had to answer for how long they already knew Wikipedia and how often they use the website.

Then, participants started the actual task: each of the eight Wikipedia articles was displayed on the screen for either five seconds or five minutes, depending on the condition. After every article, a short questionnaire about their trust in this article had to be filled in consisting of three questions: on the first question, the participants had to judge their trust in each article on a seven-point Likert scale ranging from one to seven with an ascending value indicating higher trust in the article. On the second question, they were asked to give a motivation for their evaluation in written form. In this context, there were no possible answers given, but a blank space to fill in. This method was chosen consciously to avoid social desirability bias, which describes the phenomena of participants answering questions in a way that they think they *should* answer (Thompson & Phua, 2005). In our case, social desirability bias could have the negative side-effect that a participant does not even think about "his own" features, but simply names the ones mentioned in the statement. On the third question, they were asked to rate how much they already knew about the topic. Here again, familiarity with the topic of each article was measured on a seven-point Likert scale ranging from one to seven with an ascending value indicating higher familiarity with the topic of the article. Answering all of the three questions was required before being able to continue with the study.

In total, eight Wikipedia articles were used. Of the eight articles, screenshots were made from the articles to disable the search function and links. Cues indicating the quality of the article were removed. All of them could have distracted and participants might have become confused.

Moreover, they provided an idea of how credible the article might be, a condition that would have impaired the goal of the study.

When the screenshot of the Wikipedia article appeared on the screen, it was not possible to change or edit something or to look information up on another website or search engine. While it was not possible to click on the underlined words from the Wikipedia article to follow additional links, the participants could scroll downwards to view the full article. Once the participant continued, it was not possible to return to the previous page.

Moreover, the computer was programmed in a way that the article maintained only for the exact amount of time, thus five seconds or five minutes respectively, and then disappeared automatically. It was hence not possibly for any of the respondents to have an unfair advantage through having extra time. Finally, at the end of the study, respondents had to report whether they recognized a difference in credibility and familiarity of the articles, whether they felt time pressure, whether the English language of the articles was a problem, and whether the task of the study was in line with the method they normally use when reading Wikipedia articles. At the end of the experiment, it was optional to leave remarks. The experiment took about 45 minutes.

2.4 Familiarity and quality of the articles

In this study, participants from both groups viewed a total amount of eight Wikipedia articles during the experiment. Four articles dealt with familiar topics [Angelina Jolie, an American actress; "Avatar (2009 film)", an American science fiction film and the highest grossing film of all time; Likert scale, a psychometric scale that is used in the field of psychology; and "Barney Stinson", a fictional character created for the television series "How I met your mother"], and four with unfamiliar topics ["Las Meninas" a 1656 painting by Diego Velázquez; Ed Stelmach, a Canadian politician; the Chomo Lonzo, a mountain in Tibet; and the "Kelvin probe force microscopy", a noncontact variant of atomic force microscopy (AFM)].

Familiar topics were chosen keeping the average interests of young people (music, films, and TV series) and their educational background as behavioral science students (Likert scale) in mind. Unfamiliar topics were chosen with respect to their specialty, for example a technical term from the field of microscopy [the "Kelvin probe force microscopy (KPFM)"]. In addition to that, topics that were categorized as unfamiliar were very specific, required an extraordinary general

education ("Las Meninas"), included unknown personalities for European students (Ed Stelmach) or were not the kind of subjects that make the news daily (the Chomo Lonzo).

Besides the differences in familiarity of the articles, there were differences in their quality as well: four of the Wikipedia articles were placed in a "high quality"- category ("FA", "A" and "GA"-articles) and four of them were placed in a "low quality"- category ("Start"- or "Stub"-articles). Wikipedia articles that fall under the "high quality"- category are well-written, factually accurate, neutral and contain reliable sources. Wikipedia articles falling under the "low quality"-category, on the other hand, can be incomplete, lack factual accuracy, are mostly shorter or have a small amount of references (Wikipedia Editorial Team, 2012). All articles were downloaded on June 14, 2012.

TABLE 1. Quality classes applicable for this study according to the Wikipedia Editorial Team (2012).

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 this WikiProject or elsewhere. Good article status is not a requirement for A-Class.
GA	The article has attained good article status.
Start	An article that is developing, but which is quite incomplete and may require further reliable sources.
Stub	A very basic description of the topic.

2.5 Design

The design of the study was between-subject and consequently, the study consisted of independent samples. Participants were divided randomly into the two groups. A 2 (student groups) x 2 (familiarity) x 2 (quality) design was used. Consequently, each category ("familiar/high", "familiar/low", "unfamiliar/high", and "unfamiliar/low") consisted of two articles at a time.

TABLE 2. The distribution of the Wikipedia articles used in the study with respect to their familiarity and quality.

	Article	Familiarity	Quality
1.	Angelina Jolie	Familiar	High
2.	Avatar (2009 film)	Familiar	High
3.	Likert scale	Familiar	Low
4.	Barney Stinson	Familiar	Low
5.	Las Meninas	Unfamiliar	High
6.	Ed Stelmach	Unfamiliar	High
7.	Chomo Lonzo	Unfamiliar	Low
8.	Kelvin probe force microscopy	Unfamiliar	Low

In this experiment, the articles were displayed according to the "Latin square design" used in combinatorics. In this context, a Latin square is an $n \times n$ array filled with n different symbols, each occurring exactly once in each row and exactly once in each column, and is used to determine the order the articles appear on the screen. While the numbers of the Latin square were equal to those of Table 2 [1 = Angelina Jolie, 2 = Avatar (2009 film), 3 = Likert scale, 4 = Barney Stinson, 5 = Las Meninas, 6 = Ed Stelmach, 7 = Chomo Lonzo, 8 = Kelvin probe force microscopy], the exact order of the displayed articles is shown in Table 3.

TABLE 3. The order of the displayed Wikipedia articles in the study according to the Latin square design.

2	7	8	1	4	5	6	3
1	8	3	6	7	2	5	4
8	1	6	3	2	7	4	5
7	2	1	8	5	4	3	6
4	5	2	7	6	3	8	1
3	6	5	4	1	8	7	2
6	3	4	5	8	1	2	7
5	4	7	2	3	6	1	8

2.6 Variables

The dependent variables and independent variables of the study depended on the hypotheses. In the first hypothesis, the independent variable was time, thus whether someone was part of the short time condition or part of the long time condition, and the dependent variable was the motivation for the trust judgment. In order to analyze the motivation for trust judgments, the answers of the participants on the question, where they were asked to provide a brief motivation for

their trust judgments, were coded. With 40 participants each rating eight Wikipedia articles, a total of N = 320 answers had to be coded. For this, the coding scheme based on the 3S-Model developed by Lucassen and Schraagen (2011), was used. The answers were split in four different categories: semantic features, surface features, source features, and other motivations. Semantic features are linked solely to the content of the article, its accuracy, completeness and neutrality, and contained answers such as "*I am very familiar with the topic, because I watch the TV series and so I recognized a lot of information that was correct*"[1].

Surface features are linked to features such as references, pictures, length, appearance, graphics, writing style, and internal- and external links. Source features are linked to the previous opinion and experiences of the participant with the source and contained answers such as "*I generally trust Wikipedia*"[2]. The fourth category, other motivations, included motivations that did not fit in any of the three categories.

In the second hypothesis, the independent variable was the previous experience with Wikipedia and the dependent variable was the trust in the Wikipedia articles of the study. In the third hypothesis, the independent variable was and the quality of the articles and the dependent variable was the trust in the Wikipedia articles of the study.

2.7 Data analysis

For data analysis, SPSS.18 was used for statistical analysis. To test the Inter-rater reliability, Cohen's Kappa was calculated. To test the first hypothesis, the number of motivations was counted in consideration of the category and the condition. To compare the number of the semantic features and the surface features of the short time condition with those of the long time condition, a Chi-squared test was conducted. The same was done for comparing the number of the semantic features and the surface features of the familiar articles with those of the unfamiliar articles.

To test the second hypothesis, a regression analysis was carried out to compare to mean general trust in Wikipedia with the mean trust in the Wikipedia articles of the study.

_____

[1] Translated from Dutch: "Ik ben best bekend met dit onderwerp omdat ik de serie kijk, dus ik herkende veel informatie wat waar is".

[2] Translated from Dutch: "Ik vertrouw Wikipedia op het algemeen".

To test the third hypothesis, thus to compare the trust of the high quality articles with the trust of the low quality articles both in the short time condition and in the long time condition, a Wilcoxon signed-rank test, the non-parametric alternative for the T-test for two related samples, was used.

To test the influence of time familiarity, and quality on trust; a Mann–Whitney U test was used. The Mann–Whitney *U* test is conducted for two statistically independent samples and for ordinal observations (Bertram, 2004; Jamieson, 2004). The Mann–Whitney *U* test is the non-parametric alternative for the T-test for independent samples. For some questions, a seven-point Likert scale was used and therefore the mean was calculated, as listed in section 3.1 "General results". Vos (2009), among others, points out that the calculation of Likert scores requires equal weighting. He further states that the "most common method for creating a composite score in social research simply sums the responses to items composed with Likert-style wording" (Vos, 2009; p.100). The expected score one participant could theoretically reach was thus between 1.00 and 7.00 with an ascending value indicating higher trust, higher time pressure, higher awareness of the difference in familiarity etc. However, the calculated mean score and standard deviation of Likert scales should not be seen as absolute in its validity, but more as "measure of central tendency" (Jamieson, 2004; Clegg, 1998).

3. Results

3.1 General results

The average experience of the participants with Wikipedia as a source of online information was 7.6 years (*SD = 1.82*). Of all participants, 15 percent stated that they use Wikipedia daily, 45 percent use it weekly, 32.5 percent monthly, 2.5 percent yearly, and 5 percent seldom or never. It turned out that most of them, a majority of 95 percent, never edited a Wikipedia article, whereas 5 percent did. About three quarters of the participants (77.5 percent) had used a Wikipedia article without examination to write an assignment for university, whereas 22.5 percent did not. Of all participants, 32.5 percent reported that their trust in Wikipedia seemed to be unjustified, whereas 67.5 negated this question. An analysis of the questions answered on a seven-point Likert scale, revealed that the general trust in Wikipedia was relatively high (*M = 5.18, SD = 0.87*), as was the trust in the Wikipedia articles of the study (*M = 5.08, SD = 0.57*).

Furthermore, we assume that the participants were aware of the differences in familiarity of the articles (*M = 5.63, SD = 1.58*), as well as the differences in their quality (*M = 3.90, SD =1.60*). Using a Seven-point Likert scale, the perceived time pressure was 3.88 (*SD = 2.32*). Participants from the short time condition felt extreme time pressure (*M = 6.00, SD = 0.92*), whereas those from the long time condition only felt very little time pressure (*M = 1.75, SD = 0.85*). For most of the participants, the task of the study was in line with the method that they normally use when reading Wikipedia articles (*M = 3.76, SD = 1.57*). The English language of the articles did not seem to be a problem for the participants (*M = 2.45, SD = 1.65*).

3.2 Manipulation checks

In this study, no effect of time on trust was found (*Z* = 1.55, *p* = .06), even if there was a slight difference between the trust ratings from the short time condition (*M = 4.92, SD =0.52*) and the long time condition (*M = 5.24, SD = 0.59*). The same was true for familiarity, where no effect was found on trust (*Z* = 1.15, *p* = .13). Here again, there was a slight difference in trust between the familiar (*M = 5.29, SD = 1.32*) and the unfamiliar articles (*M = 4.87, SD = 1.39*). However, the manipulation of quality turned out to be successful, as an effect of quality on trust was found (*Z* = 2.57, *p* = .01). People had more overall trust in the high-quality articles as they had in the low-quality articles, as the mean trust in the low-quality articles was 4.87 (*SD =*.70) and the mean trust in the high-quality articles was 5.29 (*SD =*.82).

TABLE 4. Trust in the information of the Wikipedia articles on seven-point Likert scales in all conditions.

	Familiar			Unfamiliar			All		
	HQ	LQ	All	HQ	LQ	All	HQ	LQ	All
Short	4.90	4.83	4.86	5.10	4.81	4.98	5.00	4.84	4.92
	(1.43)	(1.20)	(1.31)	(1.34)	(1.19)	(1.26)	(1.38)	(1.18)	(0.52)
Long	5.50	5.48	5.40	5.66	4.33	4.99	5.58	4.90	5.24
	(1.28)	(1.24)	(1.13)	(1.12)	(1.69)	(1.57)	(1.20)	(1.58)	(0.59)

HQ = high quality; LQ = low quality. Standard deviations are given in parentheses.

3.3 Hypothesis about the first impression and the use of features

Before testing the first hypothesis ("*When forming a first impression about an online article, more surface features than semantic features are applied for peoples' trust*"), the answers of the participants on the question, where they were asked to provide a brief motivation for their trust judgments, were coded as described in section 2.6.1 "Motivation for trust judgments". After the analysis of the motivations, Cohen's Kappa was calculated to determine the Inter-rater reliability. Therefore, twenty percent of all motivations ($N = 8$) were double coded by two persons. The value of Cohen's Kappa was .80, which indicates "substantial agreement" and is close to an "almost perfect agreement" according to Landis and Koch (1977). Moreover, according to Krippendorff (1980), with a Cohen's Kappa this high, a definite conclusion can be made.

In order to test the first hypothesis, the number of motivations from each category (semantic, surface, source, other) from the short time condition was compared with the number of motivations from each category from the long time condition. Checking the data about perceived time pressure described in section 3.1 "General results", it became obvious that participants from the short time condition felt time pressure, whereas participants from the long time condition did not. Testing the first hypothesis, in the short time condition, there were $N = 111$ surface features counted (69.4 percent) in comparison with $N = 12$ semantic features (7.5 percent). In the long time condition, there were $N = 91$ surface features (56.9 percent) counted in comparison with $N = 60$ semantic features (37.5 percent). As can be seen in Table 5, in both conditions, the number of surface features was higher than the number of semantic features. Moreover, further analysis revealed that the differences between the short time condition and the long time condition in the number of given certain motivations were significant, χ^2 (3, N = 320) = 52.28, $p < .01$, which supports the first hypothesis.

TABLE 5. The distribution of the different features with respect to time.

	Short time condition	Long time condition
Semantic features	12 (7.5%)	60 (37.5 %)
Surface features	111 (69.4%)	91(56.9%)
Source features	22 (13.8%)	3 (1.9%)
Other motivations	15 (9.4%)	6 (3.8%)
Total	160 (100%)	160 (100%)

Percentages are given in parentheses.

Further analysis was conducted to test the influence of familiarity on choosing certain features. For the familiar articles, there were $N = 84$ surface features (52.5 percent) counted in comparison with $N = 49$ semantic features (30.6 percent). For the unfamiliar articles, there were $N = 116$ surface features counted (72.5 percent) in comparison with $N = 24$ semantic features (15 percent). Statistical analysis revealed that the differences between familiar and unfamiliar articles were significant, χ^2 (3, N = 320) = 15.02, $p < .01$.

TABLE 6. The distribution of the different features with respect to familiarity.

	Familiar articles	Unfamiliar articles
Semantic features	49 (30.6%)	24 (15%)
Surface features	84 (52.5%)	116 (72.5%)
Source features	17 (10.6%)	11 (6.9%)
Other motivations	10 (6.3%)	9 (5.6%)
Total	160 (100%)	160 (100%)

Percentages are given in parentheses.

In total, 22.5 percent of all motivations were semantic features, 63.1 percent were surface features, 7.8 percent were source features, and 6.6 percent were other motivations. As can be seen in Table 7, the motivations for semantic features and surface features differed. In the short time condition; people relied mostly on the length (21.3 percent), images (13.8 percent), and the appearance/the lay-out (11.9 percent) when evaluating the Wikipedia article on credibility. In the long time condition, people relied mostly on references (26.3 percent), content (18.1 percent), and the length (14.4 percent) when evaluating the Wikipedia article on credibility.

23

TABLE 7. Total number of used motivations for trust in the two time conditions.

	Time condition	
	Short	Long
Semantic features		
Accuracy	0.6% (n = 1)	6.3% (n = 10)
Content	3.8% (n = 6)	18.1% (n = 29)
Completeness	1.3% (n = 2)	8.8% (n = 14)
Neutrality	0.6% (n = 1)	1.3% (n = 2)
Other semantic features	1.3% (n = 2)	3.1% (n = 5)
Surface features		
References	10.6% (n = 17)	26.3% (n = 42)
Links	1.9% (n = 3)	1.3% (n = 2)
Images	13.8% (n = 22)	3.8% (n = 6)
Length	21.3 (n = 34)	14.4% (n = 23)
Writing style	3.8% (n = 6)	1.9% (n = 3)
Appearance /Lay-out	11.9% (n = 19)	5% (n = 8)
Topic	2.5% (n = 4)	1.9% (n = 3)
Other surface features	3.8% (n = 6)	2.5% (n = 4)
Source features	13.8% (n = 22)	1.9% (n = 3)
Other motivations	9.4% (n = 15)	3.8% (n = 6)
Total	100% (n = 160)	100% (n = 160)

Numbers are given in parentheses.

3.4 Hypothesis about the previous experience

In order to test the second hypothesis ("*The previous experience with a source of online information influences people's trust*"), the general trust in Wikipedia was compared with the mean trust in the Wikipedia articles of the study. The second hypothesis was not supported, as $F (1, 39) = 3.54$ and $p = .07$. This p value still falls under the phenomenon of borderline significance (Kirkwood & Hackshaw, 2011).

Beside the complete overview of used motivations given in Table 7, an additional count was done to find reasons for the second hypothesis not being supported. Therefore, the used motivations for trust in cases where the participants scored lower in trust of a certain article than they had general trust in Wikipedia, was counted. This applied to 84 cases (of 320 in totals), 31 from the short time condition and 53 from the long time condition (see chapter 4 "Discussion" for further explanation).

TABLE 8. Number of used motivations for trust in the two time conditions in cases where the respondent scored lower on trust of a certain article than in general trust of Wikipedia.

	Time condition	
	Short	Long
Semantic features		
Accuracy	-	-
Content	-	15.1% (n = 8)
Completeness	-	7.6% (n = 4)
Neutrality	-	1.9% (n = 1)
Other semantic features	-	-
Surface features		
References	16.3% (n = 5)	28.4% (n = 15)
Links	-	1.9% (n = 1)
Images	16.3% (n = 5)	7.6% (n = 4)
Length	32.6% (n = 10)	18.9% (n = 10)
Writing style	6.5% (n = 2)	-
Appearance /Lay-out	(n = 6)	15.1% (n = 8)
Topic	6.5% (n = 2)	1.9% (n = 1)
Other surface features	3.3% (n = 1)	1.9% (n = 1)
Total	100% (n = 31)	100% (n = 53)

Numbers are given in parentheses.

3.5 Hypothesis about quality

In order to test the third hypothesis ("*There is a difference in trust between the quality of the articles in the long time condition, but not in the short time condition*"), the trust of the high quality articles was compared with the trust of the low quality articles both in the short time condition and in the long time condition. In the short time condition, the mean trust for the high-quality articles was 5.00 (*SD = 1.38*) and for the low-quality articles 4.84 (*SD = 1.18*). In the long time condition, the mean trust for the high-quality articles was 5.58 (*SD = 1.20*) and for the low-quality articles 4.90 (*SD = 1.58*). Because of the fact that there was a significant difference in credibility evaluation between the quality of the articles in the long time condition ($Z = 2.38$, $p = .02$), but not in the short time condition ($Z = 0.92$, $p = .36$); the third hypothesis is supported.
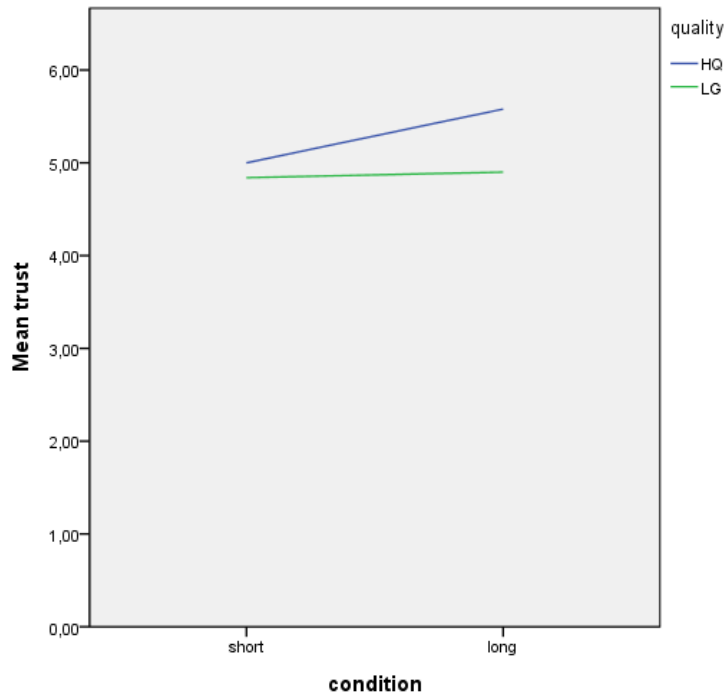
FIGURE 2. The mean trust with respect to quality and time.

4. Discussion

In this study, it was hypothesized that when people form a first impression about an online article, more surface features than semantic features are applied for their credibility evaluation. This hypothesis was supported. In addition to that, the differences between the short time condition and the long time condition in the use of semantic features and surface features were significant. Moreover, it was hypothesized that the previous experience with a source of online information influences people's rating of trust. This hypothesis was not supported.

Finally, it was hypothesized that the there is a difference in credibility evaluation between the quality of the articles in the long time condition, but not in the short time condition. This hypothesis was also supported. No effect of time and familiarity on trust was shown. However, an effect of the quality of the Wikipedia articles on trust was shown.

The fact that the first hypothesis was supported shows the "power of surface features" as it is in line with the literature from Risto (2010) and Lucassen et al. (in press), which demonstrated that people rely heavily on surface features when making a credibility evaluation of an online article. Even when people had sufficient time to read the article and to check facts, they still gave

more motivations for their trust that fall under the 'surface feature'-category. This explanation is in line with the literature that underlines the huge importance of surface features on trust, as for example the importance of such features as the design of a website (Fogg et al., 2003), a pleasing lay-out (Eysenbach & Kohler, 2002), its aesthetics (Robins et al., 2010) or the length of the article, the number of images, and the number of references (Lucassen & Schraagen, 2010; Eysenbach & Kohler, 2002). While reading the motivations from the participants, it became obvious that most of them used in fact the "classic" surface features, such as references, length, images, or the appearance and the lay-out of the text (see Table 7 for an exact overview).

The importance of semantic features, clearly shown in other studies, especially when people were familiar with the topic (Kramer, 2011; Muilwijk, 2012; Hensel, 2012), was not salient in this study. The fact that the first hypothesis was supported is also in line with findings from Lim (2009), among others, who found out that most students accept Wikipedia articles unexamined and without verifying them or checking them twice. Even when they are "forced" to rate credibility, for example in an experimental setting, they mostly base their decision on one or at most two criteria (Scholz-Crane, 1998). Metzger (2007), as well, found that only very few internet users exert themselves for evaluating the entire website and that they "scored highest on the actions that are easiest to perform and that only require their opinion" (p. 2080). As an example for such an action, she named the task to decide whether the information in an online article was complemented and current.

On the other hand, the support of the first hypothesis is in conflict with the findings from Caccioppo et al. (1986), who observed that people use the 'peripheral route' only when they are under time pressure. However, in our study, also in the long time condition, people consequently stuck to surface features and previous experience and not only when being under time pressure. However, motivation could play a role in this context: Caccioppo et al. (1986) stated that people rely on surface features when they are relatively unmotivated, which of course could have been a factor in this study, as the students did not take part voluntarily at the study but in return for credits they needed for their course of studies.

In this study, it seems that participants gained a first impression through checking the surface features like the length, the references or images; were of the opinion that this is sufficient to evaluate credibility and dispensed with reading the whole article. Even if the user is actually reading the article, he or she might only try to find confirmation for his or her first impression. In

one study, Metzger, Flanagin, and Medders (2010) found this exact behavior as participants reported to look for confirmation of their preexisting knowledge. As most of the participants stated that the task of the study was in line with the method that they normally use when reading Wikipedia articles, this would mean that they deploy the same deficient procedure of rating trust in online information in "real life". This could have negative consequences for their general knowledge or the quality of their academic studies.

The fact that the second hypothesis, which stated that the previous experience with a source of online information influences people's rating of trust, was not supported is surprisingly not in line with findings from the literature: Eysenbach and Kohler (2002) for example, stated that, for the judgment of trust, it plays an important role whether the source is an "official" and well-received one. Furthermore, Hilligoss and Rieh (2008), found out that people rely heavily on previous experiences with the source when they have to judge whether they find particular online information trustworthy or not, a conclusion also supported by Fogg (2003). Lucassen and Schraagen (2010) further underlined the importance of previous experience with the source in their 3S-model. The criterion by Lim (2009) mentioned above, that students tend to accept Wikipedia articles unexamined, does not apply in the discussion of the second hypothesis. On the contrary, it seems that, in this study, it could not be statistically established that when someone generally had found Wikipedia trustworthy, that he or she also scored high on trust on each of the eight presented Wikipedia articles and vice versa.

Even if the results fall under the phenomenon of borderline significance, reasons for the second hypothesis not being supported have to be found. One reason could be that the students were aware of the fact that, because it was about Wikipedia articles, anyone could have edited or changed the content at any time. Therefore, respondents could have been constantly aware of the fact that Wikipedia is not stable over time and not stable with regard to its quality.

Another reason could be that participants were relatively skeptic about the trustworthiness of the articles of the study, even if they have trust in Wikipedia in general. This idea arose during chats with some of the participants after the experiment when they mentioned doubts and uncertainty concerning the quality of the Wikipedia articles of the study. Many of them said things like "*I normally trust Wikipedia, but then I saw that this one article was only available in three languages and so I thought 'Oh no, that can't be trusted!'*" or "*I use Wikipedia for years, but when I saw this one article being only one page long, I concluded that they probably have*

manipulated its quality". This idea would also explain the slightly lower trust in the articles of the study than in Wikipedia in general. The long average experience of the participants with Wikipedia (7.6 years) and their frequent use of it (92.5 percent mentioned that they use the website *at least* monthly) would further support this idea.

Consequently, one could think of the prototype of a respondent not adding to support the second hypothesis, as someone who generally trusts Wikipedia, is using it for several years and is retrieving information from it frequently. However, when participating in this study; he or she recognized that, for example, the length, the appearance or the number of images or languages available, is not to be trusted and therefore rates lower on trust in the articles of the study. To further support this idea and to see what motivations people gave *exactly* when they rated a certain article of the study lower on trust than their general trust in Wikipedia, these motivations were counted (see Table 8). In this table; it is striking that, in the short time condition, people relied on mostly on the length when rating a certain article lower than their general trust; and, in the long time condition, they relied mostly on references and length.

A third reason could have to do with the fact that, during their academic studies, these students have to take part in several studies and some of them might have been similar to this one, at least when it comes to rating online information. They therefore might assume to know what the experiment is about and consequently not behave as they usually would. Sometimes during giving instructions to the participants at the beginning of the study, they were suggestive of already anticipating factual errors in the Wikipedia articles, even if there were none of these. This could be due to the fact that they were asked explicitly to "check whether they find the provided information trustworthy or not". Possibly, because of these instructions, participants were more skeptical and suspicious. For example, one girl mentioned after the experiment that she was sure that the experimenter wanted her to "*fall into the trap*" of articles manipulated in quality.

When it comes to the third hypothesis, stating that there is a difference in credibility evaluation between the quality of the articles in the long time condition, but not in the short time condition, the fact that the hypothesis was supported is in line with the literature from Risto (2010) and Lucassen et al. (in press), who demonstrated that the quality of online articles has a significant effect on trust, in the sense that people trust high-quality articles more than low-quality articles. As it was tested with time manipulation in this study, this could mean that the

respondents were under enormous time pressure and simply had no time to check the quality of the article. Therefore, there were not capable of making a final judgment about the trustworthiness of the article. In the long time condition, however, they had sufficient time to read the article and to base their credibility rating on a more extensive evaluation. The fact that the quality of the articles was not manipulated on a grand scale, even if there were differences in their quality, obviously led the participants to the conclusion that they are generally trustworthy.

When taking the findings from the literature about the first impression and its related concepts, like intuition, into account; additional support for the "dual-processing theory" was found in this study. According to this theory, people use their heuristic system, which is an implicit process resembling the concept of intuition, when they have little time for credibility evaluation and stick to their analytic system when they have more time to do so. The findings from our study suggest that, the more time people had, the higher they rated trust for the high-quality articles suggesting that they indeed used their analytic system in the long time condition to evaluate more precisely and well-thought-out. Apparently, when they had more time to read the high-quality articles, they understood that these were of a certain academic level, which they could barely recognize in the short time condition, and "rewarded" this realization with a higher trust score. On the other hand, for the low-quality articles, there was no significant difference found in trust between the two time conditions. Thus, when they had more time to read the low-quality articles, they apparently had a "feeling" for the approximate quality. Consequently, they were not convinced about their quality and the trust score hardly changed in comparison with the short time condition.

When it comes to the 3S-model of Lucassen and Schraagen (2011), the low number of other motivations (6.6 percent) in this study gives additional support to the validity of the model, as it shows that most participants indeed answered in the categories Lucassen and Schraagen proposed (semantic, surface, and source). As in the 3S-model suggested and through previous research supported, familiarity influences the motivation for the trust rating (Kramer, 2011; Muilwijk, 2012; Hensel, 2012). However, in this study, no such effect was found. There was no effect found of time on trust either. Trust was slightly higher in the long time condition, but the results were not significant.

4.1 Limitations and future research

First of all, socially desirable answers might be possible, as people were asked from the outset to give a motivation for their trust. Language could have played a role as well, as none of the participants was an English native speaker. According to Eysenbach and Kohler (2002), the trust rating is, for example, based on whether the user was capable of understanding the language. Even if most of the participants indicated that they had no difficulties in understanding the articles in English, since they were no native speaker, this could have had an influence. Another aspect could also have influenced the trust rating: four of the 20 participants from the short time condition mentioned at the end of the study that they had the impression that, for some of the articles, they had less time available than five seconds. According to them, this was due to the page loading time. As it was because of technical reasons, it was not possible to correct this during the experiment.

For future research, the number of participants could be increased, as well as the band width of their professions, meaning that it would be advisable to use another group of participants for further research besides behavioral science students. This could be done in order to have more reliable results and to avoid that participants are skeptical about the trustworthiness of the Wikipedia articles. In this context, it would be interesting to have respondents who differ in their information skills to test the influence on surface features; for example university students used to academic works and manual workers or children surfing in the internet and obtaining information just for fun and PHD-students. Finally, it would be interesting to check whether the participants, when evaluate the credibility of the articles, acted on the assumption that the quality of the articles was manipulated; and if 'yes', to which degree they believe that. This valuable additional information could be gathered through following standard think-aloud techniques; as, for example, introduced by Ericsson and Simon (1984).

References

Alexa (2012). Statistics Summary for wikipedia.org. *Retrieved, 12 May 2012, from Internet URL:* http://www.alexa.com/siteinfo/wikipedia.org?range=5y&size=large&y=t

Bertram, D. (2004). Likert Scales. Ontario, CA: University of Windsor. *Retrieved, 12 May 2012, from Internet URL:* http://poincare.matf.bg.ac.rs/~kristina//topic-dane-likert.pdf

Betsch, T. (2010). The Nature of Intuition and Its neglect in research Judgment and Decision making. In Plessner, H., Betsch, C., Betsch., T. (Eds.). Intuition in judgment and decision making. New York, NY: Taylor & Francis group, LLC.

Caccioppo, J.T., Kao, C.F., Petty, R.E., & Rodriguez, R. (1986). Central and peripheral Routes to Persuasion: An Individual Difference Perspective. *Journal of Personality and Social Psychology*, 51, 5, 1032-1043.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Clegg, F. (1998). *Simple Statistics*. Cambridge, UK: Cambridge University Press.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, 58, 737–758.

DeVellis, R.F. (2012). *Scale Development: Theory and Applications*. Los Angeles, CA: Sage Publications, Inc.

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

Evans, J., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual- process theory of reasoning. *Thinking & Reasoning*, 11(4), 382-389.

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. B*ritish Medical Journal*, 324, 573–577.

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

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. *Proceedings of the 2003 conference on Designing for user experiences*, 1-15.

Fogg, B.J., & Tseng, H. (1999). The elements of computer credibility. *CHI '99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 80-87. New York: ACM Press.

Fuchs, D. (2009). The Media's Attitude towards Wikipedia. *Retrieved, 12 May 2012, from Internet URL:* http://voices.yahoo.com/the-medias-attitude-towards-wikipedia-2847953.html?cat=15

Giles, J. (2005). Internet encyclopedias go head to head. *Nature*, 438, 900-901.

Hensel, T. (2012). Impact of duration of the search on trust judgment of Wikipedia articles. *Bachelor Thesis,* University of Twente, the Netherlands.

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing and Management*.

Hovland, C., & Weiss, W. (1952). The Influence of Source Credibility on Communication Effectiveness. *Public Opinion Quarterly*, 15, 635-650.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education,* 38, 1217-1218.

Jaschikk, S. (2007). A Stand against Wikipedia. *Retrieved, 13 May 2012, from Internet URL:* http://www.insidehighered.com/news/2007/01/26/wiki

Kahneman, D. (2003). Association Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, 93, 5, 1449-1475.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6), 515-526.

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

Kini, A., & Choobineh, J. (1998). Trust in Electronic Commerce: Definition and Theoretical Considerations. *System Sciences*, 4, 51-61.

Kirkwood, A., & Hackshaw, A. (2011). Interpreting and reporting clinical trials with results of borderline significance. *Retrieved, 25 July 2012, from Internet URL:* http://www.scribd.com/doc/79082772/Borderline-Significance

Klein, G. (1993). A recognition-primed decision (RPD) model of rapid decision making. In Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C.E. (Eds.). *Decision Making in Action: Models and Methods*. Norwood, NJ: Ablex.

Kramer, C. (2011). Personal, medium and source-dependent trust and their influence on judgment. *Bachelor Thesis,* University of Twente, the Netherlands.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159–174.

Lim, S. (2009). How and why do college students use Wikipedia. *Journal of the American Society for Information Science and Technology,* 60, 2189-2202.

Lucassen, T., Muilwijk, R., Noordzij, M.L., & Schraagen, J.M. (in press). Topic Familiarity and Information Skills in Online Credibility Evaluation. *Journal of the American Society for information Science and Technology*.

Lucassen, T., & Schraagen, J.M. (2010). Reference blindness: The influence of references on trust in Wikipedia. *ACM WebSci '11*. New York, NY: ACM.

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

McMackin, J., & Slovic, P. (2000). When does explicit justification impair decision making? *Journal of   Applied Cognitive Psychology*, 14, 527–541.

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, 2078–2091.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication* , 413-139.

Miliard, M. (2008). Wikipediots: Who Are These Devoted, Even Obsessive Contributors to Wikipedia? *Salt Lake City Weekly. Retrieved 13 May 2012, from Internet URL:* http://www.cityweekly.net/utah/article-5129-feature-wikipediots-who-are-these-devoted-even-obsessive-contributors-to-wikipedia.html

Muilwijk, R. (2012). Trust in online information. A comparison among high school students, college students and PhD students with regard to trust in Wikipedia. *Master Thesis,* University of Twente, the Netherlands.

Naegele, P.D. (2008). Who uses Wikipedia. *Retrieved, 13 May 2012, from Internet URL:* http://www.academiccommons.org/commons/essay/who-uses-wikipedia-according-powerset

Netcraft (2012). March 2012 Web Server Survey. *Retrieved, 26 August 2012,   from   Internet URL:* http://news.netcraft.com/archives/2012/03/05/march-2012-web-server-survey.html

Plessner, H., & Czenna, S. (2010). The benefits of Intuition. In Plessner, H., Betsch, C., Betsch., T. (Eds.). Intuition in judgment and decision making. New York, NY: Taylor & Francis group, LLC.

Preston, C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1-15.

Risto, M. (2010) Factors influencing trust in Wikipedia: an eye-tracking approach. *Master Thesis,* University of Twente, the Netherlands.

Robins, D., Holmes, J., & Stansbury, M. (2010). Consumer Health Information on the Web: The Relationship of Visual Design and Perceptions of Credibility. *Journal of the American Society for Information Science and Technology*, 61, 13-29.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651-665.

Scholz-Crane, A. (1998). Evaluating the future: A preliminary study of the process of how undergraduate students evaluate Web sources. *Reference Services Review,* 26(3/4), 53-60.

Smith, E., & Kosslyn, S. (2009). *Cognitive Psychology*. New Jersey, US: Pearson Prentice Hall.

Tancer, B. (2007). Look Who's Using Wikipedia. *Time Magazine*. *Retrieved, 13 May 2012, from Internet URL:* http://www.time.com/time/business/article/0,8599,1595184,00.html

Thompson, E. R., & Phua, F. T. (2005). Reliability among senior managers of the Marlowe-Crowne short-form social desirability scale. *Journal of Business and Psychology,* 19, 541-554.

Vos, H.J. (2009). *Social Research Methods*. Essex, UK: Pearson Education Limited.

Völkel M., Krötzsch M., Vrandecic D., Haller H., & Studer R. (2006). Semantic Wikipedia. *Retrieved, 13 May 2012, from Internet URL:* http://www.aifb.uni-karlsruhe.de/WBS/hha/papers/SemanticWikipedia.pdf

Walraven, A. (2008). Becoming a critical websearcher – Effects of instruction to foster transfer. *Retrieved, 11 May 2012, from Internet URL:* http://www.ou.nl/documents/Promoties-en-oraties/Promoties/Promoties2008/Proefschrift_Amber_Walraven_2008.pdf

Wannemacher, K. (2009). Articles as Assignments – Modalities and Experiences of Wikipedia Use in University Courses. Advances in Web Based Learning – ICWL 2009. *Lecture Notes in Computer Science*, 5686, 434-443

Wathen, N. C., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology*, 53(2), 134-144.

Wikipedia (2012). Size of Wikipedia. *Retrieved, 12 May 2012, from Internet URL:* http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Wikipedia Editorial Team (2012). Wikipedia: Version 1.0 Editorial Team. *Retrieved, 10 June 2012, from Internet URL:* http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

Woodson, A. (2007). Wikipedia remains go-to site for online news. *Retrieved, 12 May 2012, from Internet URL:* http://www.reuters.com/article/internetNews/idUSN0819429120070708.