Reference Blindness:
The Influence of References on Trust in Wikipedia

Teun Lucassen
t.lucassen@utwente.nl

Matthijs L. Noordzij
m.l.noordzij@utwente.nl

Jan Maarten Schraagen
j.m.c.schraagen@utwente.nl

Department of Cognitive Psychology & Ergonomics
University of Twente
P.O. Box 215, 7500 AE, Enschede
The Netherlands

ABSTRACT

In this study we show the influence of references on trust in information. We changed the contents of reference lists of Wikipedia articles in such a way that the new references were no longer in any sense related to the topic of the article. Furthermore, the length of the reference list was varied. College students were asked to evaluate the credibility of these articles. Only 6 out of 23 students noticed the manipulation of the references; 9 out of 23 students noticed the variations in length. These numbers are remarkably low, as 17 students indicated they considered references an important indicator of credibility. The findings suggest a highly heuristic manner of credibility evaluation. Systematic evaluation behavior was also observed in the experiment, but only of participants with low trust in Wikipedia in general.

1. INTRODUCTION

The introduction of Wikipedia in 2001 has sparked a lot of discussion. Many researchers question how an encyclopedia can ever be a credible source of information, when anyone can change its contents [3, 13]. Nevertheless, the quality of the articles has been proven to be quite high in comparison to professionally maintained databases [12]. It has even been shown that the accuracy of Wikipedia is similar to a traditional encyclopedia[6]. However, due to the open editing model, the risk of encountering false information is always present [2]. Therefore users should always assess the credibility of the presented information.

Some confusion exists in literature about credibility and trust, terms which are often used interchangeably [5]. In this paper, we consider credibility a property of the information, which can be explained as believability. Based on this property, users may decide to trust or not trust the information. This decision always involves a degree of risk, as users can never by entirely certain about the credibility of information. In order to reduce this risk, credibility evaluations can be performed, in which several aspects of the information are used as indicators of credibility. These aspects vary between different situations and users [4]; examples are text length, writing style, or images [9].

The extent to which credibility is actually being evaluated by users, is heavily dependent on the context of the information. To explain these differences, dual-processing theory [1] can be helpful. It has been proposed that credibility evaluation can be carried out in a peripheral/heuristic or central/systematic manner [5, 7, 11]. The choice between these may for instance depend on motivation (e.g., consequences of poor information), ability (e.g., information skills), purpose of the information (e.g., school assignment) and familiarity with the topic.

Figure 1: The relationship between systematic and heuristic evaluation and the corresponding features of references.
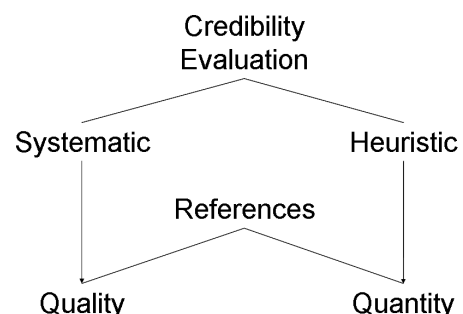
It has been shown that college students are capable of performing meaningful credibility evaluations of Wikipedia articles [9]. In a think-aloud experiment, students could successfully distinguish high quality articles from low quality articles, even when they were unfamiliar with the topic at hand. Protocol analysis has revealed that their evaluations were to a large extent based on the quality and quantity of the references in the articles (covering 26% of all utterances).

Assessing the quality of references can be seen as systematic evaluation, as this requires effortful processing of the reference list, deciding for each entry whether it is a credible and relevant source. In contrast, the evaluation of the quantity

of references (length of the reference list), is highly heuristic behavior, which can be performed in one single glance. This relationship is illustrated in Figure 1.

In this paper we investigate whether college students truly evaluate references both heuristically and systematically. We do this by manipulating the quality and quantity of the references of Wikipedia articles, corresponding with respectively systematic and heuristic evaluation. Given the importance of references for credibility evaluations of college students as suggested in [9], this leads to the following two hypotheses:

HYPOTHESIS 1. *Reference quality has a positive impact on information credibility when a systematic evaluation is performed.*

HYPOTHESIS 2. *Reference quantity has a positive impact on information credibility when a heuristic evaluation is performed.*

In the discussion of heuristic versus systematic credibility evaluation, it is assumed that an active evaluation of the information at hand is performed. However, an alternative strategy has been proposed in the 3S-model [10]. Instead of considering various features from the information (heuristically or systematically), one may also consider the source of the information. This may impose a strong bias on the evaluation as positive and negative prior experiences may lead to respectively high and low trust, without considering the information itself.

Consider for instance a user who has a lot of positive experiences with a particular source. When this user encounters new information from the same source, he or she may not feel the need for a thorough, systematic evaluation and may thus only perform a quick, heuristic evaluation of credibility. On the other hand, when a user has low trust in a source due to negative prior experiences, he or she is likely to be very cautious about the information, resulting in a systematic evaluation. This leads to the third and fourth hypotheses:

HYPOTHESIS 3. *Users with high trust in the source perform a heuristic credibility evaluation.*

HYPOTHESIS 4. *Users with low trust in the source perform a systematic credibility evaluation.*

These hypotheses can also be explained by the dual processing model of website credibility evaluation [11]. In this model, the choice between a heuristic or systematic evaluation depends on the motivation and ability of the user. As suggested, positive or negative prior experiences may influence the motivation to evaluate. The other factor, ability, is assumed to be constant in this study, as we consider college students, who have been shown to be able to evaluate credibility [9].

2. METHOD

2.1 Participants
A total of 23 college students (7 male, 16 female) participated in the experiment. Their ages varied between 18 and 24 years ($M = 19.9, SD = 1.9$). All participants were Dutch ($N = 10$) or German ($N = 13$). Course credits were awarded after attendance.

2.2 Task
The participants in this experiment were asked to perform the Wikipedia Screening Task [9]. In this task, a Wikipedia article is presented, in which any direct cues about its credibility (such as *[citation needed]* remarks) are removed. The participant has to indicate how much trust he or she has in the article, along with a rationale for their answer.

Each participant viewed the same four articles obtained from the English Wikipedia. The topics used were "Comet", "Infrared", "Linux Kernel", and "Volcano". All articles were of average (B-Class) quality as rated by the Wikipedia Editorial Team[1] and assumed to be of similar familiarity for our participants.

2.3 Design
A 2×2 repeated measures design was employed. Quality and quantity of references were varied within-subject.

The quality of the references was manipulated by replacing the original references by those of different, completely unrelated articles. Table 1 shows the topics of the references for each article in the low quality condition.

Table 1: Topics of the references in the low quality condition

Article topic	Reference topic
Comet	Pope
Infrared	Triceratops
Linux Kernel	Stem cell
Volcano	Money

The quantity of the references was manipulated by adjusting the number of references, resulting in two conditions: short (about 5 references) and long (about 25 references).

The conditions were presented in the following fixed order: (1) high quality-long, (2) high quality-short, (3) low quality-long, and (4) low quality-short. Four versions of each article were created to match each of the four conditions. The order of articles was balanced over the participants using a Latin square design.

2.4 Procedure
Upon arrival, participants signed a consent form and provided demographical information and an indication of their general trust in Wikipedia (on a 7-point Likert scale). After this, they were instructed on the Wikipedia Screening Task. One practice article (on the topic "Fruit") was presented to familiarize the participants with the task. Subsequently, the actual experiment started. After each article, a

---

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

questionnaire was provided, on which the participants rated credibility on a 7-point Likert scale. Additionally, they were asked to provide a rationale for their answers. No time limit was set for each article. After the participants evaluated all four articles, they were asked whether they considered references in their credibility assessments and whether they noticed the manipulations of the references. The experiment took about 30 minutes.

3. RESULTS
Table 2 shows the average trust (on 7-point Likert scales) of the participants in each condition.

Table 2: Average trust in each condition (standard deviation between parentheses)

Condition	Trust
high quality	5.72 (1.13)
low quality	4.80 (1.72)
long reference list	5.46 (1.49)
short reference list	5.07 (1.54)

Articles with references of high quality were trusted more by the participants than articles with low-quality references ($t(22) = 3.07, p = .003$), indicating that systematic evaluations were performed during the experiment. Articles with long reference lists were trusted more than articles with short reference lists ($t(22) = 2.05, p = .027$), indicating that heuristic evaluations were also performed during the experiment.

A median split was performed on general trust in Wikipedia. For participants with low general trust, the quality manipulation had a negative effect on trust ($t(9) = 2.85, p = .01$), whereas no effect of the quantity manipulation was found ($t(9) = 1.00, p = 0.17$). For participants with high general trust, the effect was the reverse. The quality manipulation had no effect on trust ($t(9) = 1.69, p = .058$), whereas the quantity manipulation influenced trust ($t(9) = 1.80, p = 0.048$). This supports our hypotheses that a systematic evaluation is performed when general trust is low and a heuristic evaluation is performed when general trust is high, although we acknowledge that the differences in trust of participants with high general trust are relatively small.

Interestingly, in the questionnaire after the experiment, 74% of the participants (17 of 23) indicated to have paid attention to the references in their credibility assessments. However, only 26% of the participants (6 of 23) had noticed that in half of the presented articles, the references were not related to the topic of the article. Furthermore, 39% of the participants (9 of 23) had noticed the differences in length.

4. DISCUSSION
The experiment in this study has revealed novel insights in the use of references in credibility evaluation.

First of all, the quality of the references had a positive influence on trust in the information, providing support for our hypothesis that systematic evaluations are performed. However, further analysis showed that this was only the case when trust in the source (Wikipedia) was low. This

supports our hypothesis that low trust in the source leads to systematic credibility evaluation.

Length of the reference list also influenced trust for the participants in our experiment. This supports our hypothesis that heuristic evaluations are performed. Furthermore, it was shown that only when general trust was high, that length was influential. This in turn supports our hypothesis that high trust in the source leads to the use of heuristics in credibility evaluation.

Although both quality and quantity influenced trust in the information, it was seen that the effect size was much larger for quality. One could derive from this that reference quality is more important than the number of references. However, an alternative explanation lies in the extent to which both variables were manipulated. Whereas in the low-quality condition, we assured that the references were not of any relevance to the topic, and thus of no quality at all, in the low-quantity condition, the articles still featured about 5 references. This number may have been sufficiently high for a number of participants to still evaluate the article as being credible. It is also possible that the number of references is considered dichotomously, and that the presence of any number of references (or at least five) is considered sufficient for the credibility of an article.

Perhaps the most remarkable observation is that only 6 of the 23 participants noticed that the references were not related to the topic of the article in the low-quality condition. However, 17 participants indicated to have paid attention to the references. We coin this phenomenon *reference blindness*: users consider references important for credibility, but as long as they are present, the quality of the references mostly does not seem to matter. Only when users are suspicious of the source of the information and thus perform a thorough, systematic evaluation, the quality of the references influences trust. Otherwise, heuristic evaluation is the dominant strategy, even when users are specifically asked to evaluate credibility as was the case in this study.

5. FUTURE RESEARCH
This study has indicated the complex nature of the use of references in credibility evaluation. More research could shed more light on the phenomenon of *reference blindness*. A promising method in future research is eye-tracking, as this gains insight in the visual attention of users performing this task. It would be very interesting to see how much attention is paid to the references.

Furthermore, a convenience sample consisting of college students was used in our experiment. While college students are an important group of users on Wikipedia[8], an academic bias can be expected concerning the importance of references. Other populations with different characteristics (e.g., education level, age) should also be considered, for instance high school students or even younger children.

Finally, references are only one of the features from the information that can be used in credibility evaluation. Other features (such as text length, images, or writing style) could also be systematically manipulated through an experiment, investigating their importance in credibility evaluation.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] S. Chaiken and D. Maheswaran. Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgment. *Journal of Personality and Social Psychology*, 66(3):460–473, Mar. 1994.

[2] P. Denning, J. Horning, D. Parnas, and L. Weinstein. Wikipedia risks. *Commun. ACM*, 48(12):152, Dec. 2005.

[3] P. L. Dooley. Wikipedia and the two-faced professoriate. In *WikiSym '10: Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, pages 1–2, New York, NY, USA, 2010. ACM.

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

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

[6] J. Giles. Internet Encyclopaedias go head to head. *Nature*, 438(7070):900–901, Dec. 2005.

[7] B. Hilligoss and S. Rieh. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, 44(4):1467–1484, July 2008.

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

[9] T. Lucassen and J. M. Schraagen. Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*, WICOW '10, pages 19–26, New York, NY, USA, 2010. ACM.

[10] T. Lucassen and J. M. Schraagen. Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology*, in press.

[11] M. J. Metzger. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *J. Am. Soc. Inf. Sci.*, 58(13):2078–2091, Nov. 2007.

[12] M. S. Rajagopalan, V. Khanna, M. Stott, Y. Leiter, T. N. Showalter, A. Dicker, and Y. R. Lawrence. Accuracy of cancer information on the Internet: A comparison of a Wiki with a professionally maintained database. *J Clin Oncol (Meeting Abstracts)*, 28(15_suppl):6058+, May 2010.

[13] N. L. Waters. Why you can't cite Wikipedia in my class. *Commun. ACM*, 50(9):15–17, Sept. 2007.