Topic Familiarity and Information Skills in Online Credibility Evaluation

Teun Lucassen, Rienco Muilwijk, Matthijs L. Noordzij, and Jan Maarten Schraagen
Department of Cognitive Psychology and Ergonomics, University of Twente, P.O. Box 215, 7500 AE Enschede, The Netherlands. E-mail: {t.lucassen, m.l.noordzij, j.m.c.schraagen}@utwente.nl, riencomuilwijk@gmail.com

With the rise of user-generated content, evaluating the credibility of information has become increasingly important. It is already known that various user characteristics influence the way credibility evaluation is performed. Domain experts on the topic at hand primarily focus on semantic features of information (e.g., factual accuracy), whereas novices focus more on surface features (e.g., length of a text). In this study, we further explore two key influences on credibility evaluation: topic familiarity and information skills. Participants with varying expected levels of information skills (i.e., high school students, undergraduates, and postgraduates) evaluated Wikipedia articles of varying quality on familiar and unfamiliar topics while thinking aloud. When familiar with the topic, participants indeed focused primarily on semantic features of the information, whereas participants unfamiliar with the topic paid more attention to surface features. The utilization of surface features increased with information skills. Moreover, participants with better information skills calibrated their trust against the quality of the information, whereas trust of participants with poorer information skills did not. This study confirms the enabling character of domain expertise and information skills in credibility evaluation as predicted by the updated 3S-model of credibility evaluation.

Introduction

Nowadays, we live in a world in which anyone can go online to attain all information imaginable, and more. Although this presents the opportunity to expand our knowledge very quickly, the freedom of the Internet also has its downside. One particular issue is that of the credibility of online information. In the pre-Internet era, evaluating credibility was relatively easy, as usually a specific individual could be held accountable (i.e., the author). Moreover, this task was mostly performed by trained professionals, such as newspaper or book editors. Nowadays, credibility evaluation is increasingly a responsibility of the end user, who often lacks the required skills (and motivation) for the job (Flanagin & Metzger, 2007). The second wave of Internet technology (Web 2.0) has amplified this problem, because nowadays anyone can make information available to everyone.

The topic of credibility evaluation in online environments has attracted numerous researchers trying to explain the behavior of Internet users. The influence of many aspects, such as user characteristics (Hilligoss & Rieh, 2008; Metzger, 2007), information features (Lucassen & Schraagen, 2010; Yaari, Baruchson-Arbib, & Bar-Ilan, 2011), or other situational factors (Fogg, 2003; Kelton, Fleischmann, & Wallace, 2008) have been shown. One particular study demonstrated the impact of three distinctive user characteristics (domain expertise, information skills, and source experience) on the information features used in credibility evaluation (Lucassen & Schraagen, 2011). Initial validation for the proposed relationship between user characteristics and information features (in the 3S-model, explained below) was provided by means of an online quasi-experiment, which mainly focused on the influence of domain expertise.

In the current study, we attempt to gain more insight into the influence of various user characteristics on credibility evaluation. Two key user characteristics for active credibility evaluation (domain expertise and information skills) are manipulated and controlled systematically in a think-aloud experiment, in order to better understand their relationship with credibility evaluation and, ultimately, trust. Moreover, the experiment can show which particular strategies to evaluate credibility are applied by various users.

The remainder of this article is structured as follows. We start by discussing and defining the concepts of trust and credibility in online environments. After this, the 3S-model (Lucassen & Schraagen, 2011) is discussed and revised, and related research is reviewed. Our method to explore the role of domain expertise and information skills in credibility evaluation is explained, followed by the results. The article ends with a discussion of the results and their implications for academic research and practice.

Received April 16, 2012; revised May 12, 2012; accepted May 14, 2012

 DOI: 10.1002/asi.22743

Trust and Credibility Evaluation

Trust is an important concept in a world where we rely on interactions with other people (e.g., financial transactions, information exchange). Constant monitoring of the other person is often impossible, so we need to have trust in this person such that his or her actions are beneficial (or at least not detrimental) to us (Mayer, Davis, & Schoorman, 1995). This implies that a certain risk is taken each time we trust someone (Kelton et al., 2008).

Trust in other people during information exchanges, trust can be defined as the expectation that the information is correct. The aspect of information on which users base their trust is called credibility. Hence, trust can be seen as a property of the user, whereas credibility is a property of information (Lucassen & Schraagen, 2011). In psychology, two key elements of credibility are defined: trustworthiness and expertise (Fogg & Tseng, 1999). The first refers to whether someone *wants* to give correct information (well-intentioned), whereas the latter refers to whether they are *able* to do so (knowledgeable). Information usually travels from one person to another, so one-way relationships between the reader and author can be expected rather than mutual relationships (Kelton et al., 2008).

In some situations, people may want to reduce the risk they take when they trust information (or similarly: trust that the other person gives us correct information). This may, for instance, be the case when the consequences of incorrect information are high (i.e., making important, but wrong, decisions based on the information). The risk of trusting can be reduced by performing a credibility evaluation. In such an evaluation, users search for cues in the information that they apply as indicators of high or low credibility. Which cues these are is largely dependent on the mental model of trust of each individual user (Hilligoss & Rieh, 2008). Different users may have very different conceptions of what is important for credibility. It has, for instance, been shown that references are a very important indicator of credibility for college students (Lucassen, Noordzij, & Schraagen, 2011). This can be explained by an academic bias towards references. Users without academic training are expected to attribute less value to this particular cue. They may pay more attention to other aspects, such as understandability or images.

The extent to which a credibility evaluation is performed is dependent on the motivation and ability of the user (Metzger, 2007). Following dual-process theory (Chaiken, 1980), users only perform a credibility evaluation when they have a motivation to do so. According to Metzger (2007), motivation stems from the "consequentiality of receiving low-quality, unreliable, or inaccurate information" (p. 2,087). Moreover, the level of processing (i.e., heuristic vs. systematic) is dependent on the ability (skills) of the user; a systematic evaluation is thus only performed when a user is motivated and able to evaluate.

It should be noted that the apparent dichotomous choice between heuristic and systematic processing is somewhat simplistic in the domain of trust. Credibility evaluation as a strategy to reduce the risk of trusting is always heuristic to a certain extent. This claim can be illustrated by considering the extreme case of systematic processing. If a user would consider all aspects of credibility systematically, she or he would be certain of the credibility of the information. This means that the concept of trust is eliminated. Hence, absolute systematic processing is not possible in credibility evaluation, and therefore always remains heuristic to a certain extent.

The 3S-Model

In order to better understand how people form their judgments on the credibility of information, the 3S-model was introduced by Lucassen and Schraagen (2011). In this model (see Appendix), three strategies of credibility evaluation are proposed.

The first strategy is to consider semantic features of the information, such as its accuracy or neutrality. This requires a certain level of domain expertise from the user, as the presented information is compared with his or her own knowledge on the topic. Following this strategy, the most salient aspect of credibility is addressed: factual accuracy.

When domain expertise is low or nonexistent, it is nearly impossible to follow the semantic strategy. Users can work around this deficit by considering surface features of the information. These features pertain to the way the information is presented. It has, for instance, been shown that the design of a website is one of the most important indicators of credibility (Fogg et al., 2003). Moreover, the aesthetics of a website haven been shown to correlate with perceived credibility on multiple occasions (Robins & Holmes, 2008; Robins, Holmes, & Stansbury, 2010), with beautiful designs being judged more credible. Considering the layout of Wikipedia articles as an indicator for credibility is less useful, as all articles have the same look and feel. Salient indicators in this context are, for instance, the length of the article, the number of references, and the number of images (Lucassen & Schraagen, 2010). The strategy of considering surface features requires different skills from the user, namely, generic information skills. Such skills include knowledge of how particular features are related to the concept of credibility (e.g., the presence of references suggests well-researched information).

A third strategy is to consider previous experiences with a particular source as an indicator of credibility. As opposed to the first and second strategy, this is a passive strategy, as the actual information itself is not considered, but only the source where it came from.

When evaluating credibility, users follow one or more of these strategies, leading to a trust judgment. Following the outcome of the trust judgment, source experience can be adjusted accordingly.

In this study, we propose a slightly revised version of the 3S-model (Figure 1). Conceptually, the model remains

DOI: 10.1002/asi

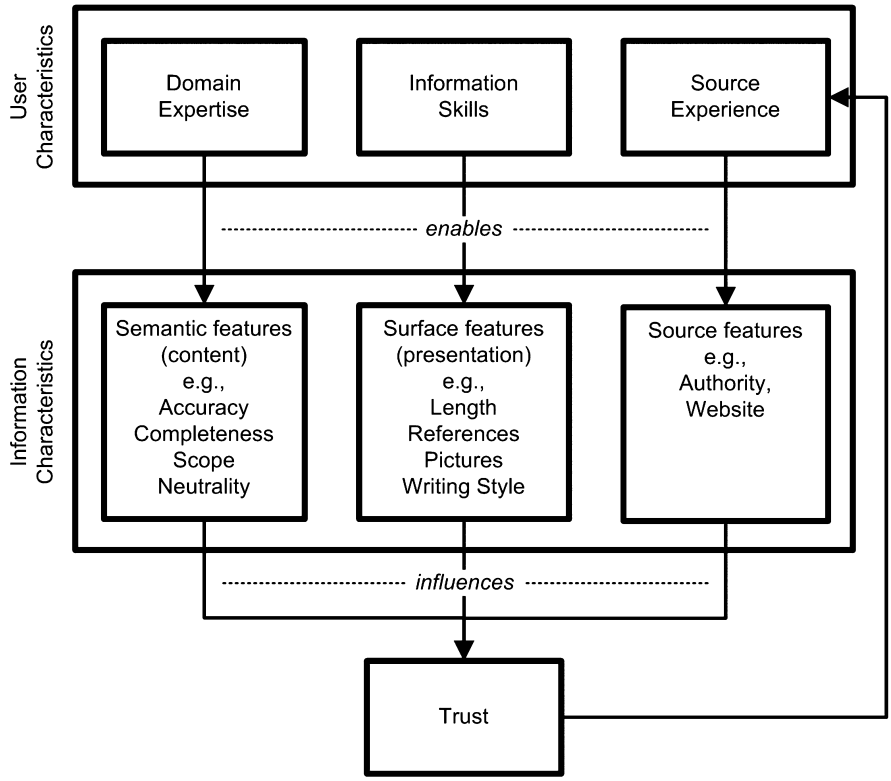

FIG. 1. Revised 3S-model of credibility evaluation.

unchanged, but two minor adjustments have been made to enhance the clarity of the visualization.

First, the original term "Trust Judgment" proved to be ambiguous. It can be interpreted as the *process* of judging trust (i.e., credibility evaluation, considering the various related features) or as the *outcome* of this process (i.e., trust in the information). Because the 3S-model is more of an information model than a process model, we decided to rename "Trust Judgment" as "Trust" in the revised version.

Moreover, in the original model the connecting arrows only indicated that a relationship existed between certain user characteristics and information features. By switching the position of information characteristics and user characteristics in the original model, we are better able to specify the nature of these relationships. The three user characteristics play an enabling role in the selection of information features; for instance, possessing domain expertise on a topic enables the utilization of semantic features. The same goes for information skills and source experience: possessing information skills enables the utilization of surface features and possessing source experience enables the utilization of source features in credibility evaluation. Considering the enabling character of the user characteristics, it naturally follows that only those information characteristics that are enabled have an influence on trust. Consider, for example, a college student with no particular knowledge of the topic at hand. We can expect a reasonable level of information skills, which she or he can bring to bear when evaluating credibility. However, his or her domain expertise on the topic is low or even nonexistent. This means that when the information looks credible on the surface level (e.g., lengthy, numerous references and images), factual errors in the information are likely to go undetected by this student. This may result in (unjustifiably) high trust in the information.

It is perhaps tempting to interpret the semantic strategy as a systematic approach and the surface strategy as a heuristic approach. However, this is not necessarily the case. For example, recognizing stated facts as you have learned them before can be considered heuristic processing (Klein, Calderwood, & Clinton-Cirocco, 1986), but classifies as a semantic strategy. Similarly, considering the quality of each of the references is clearly systematic processing (Lucassen et al., 2011), but classifies as a surface strategy. Hence, both the semantic and surface strategy can be performed systematically and heuristically. An exception has to be made for the source strategy: this is largely heuristic, as the actual information is not considered at all, only where it came from. Earlier experiences with this source are in this case a predictor of the credibility of the current information.

Domain Expertise

The important role of domain expertise in credibility evaluation has been shown on numerous occasions. It has been demonstrated that having knowledge on the topic at

DOI: 10.1002/asi

hand leads to more trust (Chesney, 2006; Self, 1996) in the information. However, we argue that this is not necessarily the case. Following Lucassen and Schraagen (2011), trust will only be high if the features incorporated in the evaluation indicate high credibility. Hence, domain experts will only have high trust in information that is credible at the semantic level (e.g., factually accurate). This claim is supported by Chesney (2006), who argued that Wikipedia is credible, since domain experts trusted the information more than novices.

Domain experts are expected to evaluate credibility better than novices. Kelton et al. (2008) argued that their trust is better calibrated to the actual credibility of information, as their general propensity to trust has less influence on their judgments than novices' propensity to trust.

Information Skills

Information skills, or information literacy, can be defined as "the skills required to identify information sources, access information, evaluate it, and use it effectively, efficiently, and ethically" (Julien & Barker, 2009, p. 12). Brand-Gruwel, Wopereis, and Vermetten (2005) defined five stages for information problem solving, namely defining, selecting, searching, processing, and organizing. Information skills influence how well information users perform each of these tasks. It has, for instance, been shown that experts (PhD students) spend more time defining the problem than novices (undergraduates) before moving on to subsequent stages (Brand-Gruwel et al., 2005). Moreover, information experts more often activate their prior knowledge, elaborate on the content, and regulate their process. These differences result in a better task performance by experts.

Based on the definition given above, information skills do not only relate to the ability to evaluate credibility. However, it is an important sub-skill, which relates to multiple stages of the information problem-solving process. For instance, source credibility is of importance when selecting appropriate information sources, but when processing information that was found, credibility at the surface and semantic level can be evaluated.

According to Alexander and Tate (1999), users who evaluate information should focus on the following five criteria: accuracy, authority, objectivity, currency, and coverage. Walraven, Brand-Gruwel, and Boshuizen (2009) showed that students often know more of such criteria than they actually apply when searching for information, indicating that they lack a critical disposition to information from online sources. Moreover, Julien and Barker (2009) demonstrated large gaps in the information skills of students (e.g., lack of knowledge on how search engines work), arguing that education in information skills should be improved, especially at high schools.

Following the various studies on information skills of students at various educational levels (i.e., high school, undergraduate, postgraduate) it can be concluded that information skills improve with education. High school students have very limited skills to evaluate information, which means that they largely depend on the credibility of a source (e.g., university websites are credible) rather than evaluating the content itself (Julien & Barker, 2009). Undergraduate students are better able to evaluate information, largely by applying various heuristics (on the source of information and the content itself; Hilligoss & Rieh, 2008). Postgraduate (PhD) students can be considered experts in information problem solving (at least in comparison with undergraduate students), as they focus much more on various aspects (e.g., quality, relevance, reliability) of the actual content of information (Brand-Gruwel et al., 2005).

Wikipedia

One particularly interesting source on the web to study the evaluation behavior of lay users is Wikipedia. This vast online encyclopedia thrives on user contribution: everyone can make changes or additions to the available articles, or create new ones. Intuitively, this seems like a bad idea, as this open-editing model is bound to attract vandals (Viégas, Wattenberg, & Kushal, 2004) and other individuals with bad intentions (consider the "trustworthiness" aspect of credibility). Moreover, how can we know that contributors have the appropriate credentials (consider the "expertise" aspect of credibility; Fogg & Tseng, 1999) to add information?

Still, history has proven many of the early critics wrong. Wikipedia has been shown to be a reliable source of information on numerous occasions (e.g., Chesney, 2006; Giles, 2005; Rajagopalan et al., 2010). This has been attributed to the collaborative manner in which the articles are written (Wilkinson & Huberman, 2007). However, because of this very same principle, Wikipedia users can never be entirely certain of the credibility of the articles. This creates the need for trust and thus also the need to evaluate credibility before using the information.

The notion that traditional heuristics no longer apply on the web is also true in the domain of Wikipedia, perhaps to an even larger extent due to its open-editing model (Magnus, 2009). This of course has implications for credibility evaluation. Lucassen and Schraagen (2011) showed that when factual errors are present in Wikipedia articles, trust is only influenced when the user is a domain expert, and even then only to a limited extent. Novices were not influenced at all by the factual errors. In an earlier study (Lucassen & Schraagen, 2010), undergraduate students worked around their lack of domain expertise by applying their information skills. By doing so, they were able to distinguish between high- and low-quality information on Wikipedia (Lucassen & Schraagen, 2010).

Hypotheses

In the original study on the 3S-model (Lucassen & Schraagen, 2011), initial validation through an online quasi-experiment was provided. This was done by manipulating a

DOI: 10.1002/asi

key semantic feature (factual accuracy) and showing that users with some domain expertise were influenced by errors, whereas complete novices were not. However, this approach has certain limitations, which we address in this study.

The participants in the preceding study were recruited on the basis of their domain expertise (high vs. low) in the field of automotive engineering. This means that their level of information skills was not controlled for. The first goal in this study was to further explore the influence of both domain expertise and information skills on the features used in credibility evaluation. Domain expertise was manipulated in a more rigorous fashion by presenting the participants information on topics on which they indicated themselves to have high or low prior knowledge. In line with the results of Lucassen and Schraagen (2011), we formulate the following hypothesis:

Hypothesis 1: Users with more domain expertise utilize more semantic features in credibility evaluation than users with less domain expertise.

Information skills were controlled by selecting three different groups that are known to differ in their level of information skills: high school students, undergraduates, and postgraduates (Brand-Gruwel et al., 2005; Julien & Barker, 2009). Naturally, these groups will also differ on other dimensions than information skills only (e.g., age), which may introduce confounding variables in our experiments. However, these three groups are all regular information seekers in comparable contexts (i.e., education), which adds to the external validity of this research. In contrast, we believe that a more controlled but isolated approach (e.g., training one half of a coherent group with low information skills) would harm the validity of this study.

Following the preceding discussion on information skills, we expect that users with better information skills (e.g., postgraduate students) can bring to bear more strategies to consider various features of the information, rather than only focusing on the factual accuracy (i.e., semantics features) of information. By doing so, they can work around their lack of prior knowledge by considering surface features. In contrast, users with poorer information skills will incorporate fewer surface features, as they are unfamiliar with such indicators of credibility. Instead, they will mainly consider the semantics of the information, also when they have limited prior knowledge on the topic at hand. This leads to the second hypothesis:

Hypothesis 2: Users with better information skills utilize more surface features in credibility evaluation than users with poorer information skills.

As noted, in the original experiment a semantic feature was manipulated. This means that differences in trust between users were mostly caused by differences in their domain expertise. The application of surface features in credibility evaluation was also shown in the experiment. However, the articles were kept unchanged on the surface level, which means that although information skills were applied, this had no influence on trust.

TABLE 1. Key characteristics of all participant groups.

			Gender		Nationality	
	N	Age	Male	Female	Dutch	German
High school	13	14.3 (0.6)	5	8	13	0
Undergraduate*	12	23.4 (6.3)	5	7	7	5
Postgraduate	15	27.0 (1.9)	7	8	13	2

Note. Standard deviation for age is given in parentheses.
*For the undergraduate group, the uncoded transcriptions (raw utterances of the participants of Lucassen and Schraagen [2010] were used in the data analyses.

In this study, we manipulate the quality of the presented information following the classification of the Wikipedia Editorial Team ("Wikipedia:Version 1.0 Editorial Team," n.d.). Their goal is to assess all Wikipedia articles on how close they are to a "distribution-quality article on a particular topic." Although this implies that the articles should be factually accurate, we expect that the difference between high-quality articles and low-quality articles is best visible on the surface level, for instance by the number of references, its length, and the presence of images. These characteristics are explicitly noted in the grading scheme (see also Table 2) of the Wikipedia Editorial Team.

When the quality levels of the Wikipedia Editorial Team are indeed best visible at the surface level, this means that a certain level of information skills is needed in order to be influenced by the quality. We expect that users with poorer information skills do not focus on the features that reflect the quality level, and are thus not influenced by them. This leads to the following hypotheses:

Hypothesis 3: Trust of users with better information skills is influenced by the quality of the information.

Hypothesis 4: Trust of users with poorer information skills is not influenced by the quality of the information.

In contrast, we do not expect that domain expertise has much influence on trust in high-quality or low-quality information. Articles with lower quality are generally also not expected to feature major errors; they are mainly much shorter and unfinished compared to higher-quality articles.

Method

Participants

A total of 40 participants took part in the experiment. Three participant groups were created: high school students, undergraduate students, and postgraduate (PhD) students. Table 1 shows the key characteristics of each of the participant groups.

The high school students were in their third year out of 6 years of preacademic education (i.e., preparing them for a subsequent university or college education). They received monetary compensation for their participation. Their experience with Wikipedia ranged from 2 to 5 years with an

DOI: 10.1002/asi

average of 4. Only three of the high school students mentioned the open-editing model behind Wikipedia when asked to explain the basics of this website. One high school student had experience in editing articles on Wikipedia himself.

The undergraduates were all following education in the domain of behavioral sciences. They received course credits for participating. Their experience with Wikipedia ranged from 3 to 8 years with an average of 5. All undergraduates students were able to explain the basics of Wikipedia in their own words. None of them had contributed to Wikipedia before.

The postgraduates were from various disciplines, such as behavioral sciences, physics, and management sciences. Their experience with Wikipedia ranged from 4 to 10 years with an average of 7. All postgraduates described the online encyclopedia as an open source that anyone can edit. Three postgraduates had experience in editing articles on Wikipedia.

All participants in the three groups were proficient in the Dutch language and able to effortlessly express their thoughts in this language. Therefore, Dutch was chosen for the think-aloud method (Ericsson & Simon, 1984). The articles used in the experiment were obtained from the English Wikipedia for the undergraduate and postgraduate students. No major language barriers were reported after the experiment. The participating high school students were not sufficiently proficient in the English language to be able to fully comprehend information in this language, therefore the Dutch Wikipedia was used to select articles for this participant group.

Task

The participants performed the Wikipedia Screening Task (Lucassen & Schraagen, 2010). In this task, a Wikipedia article is displayed in a web browser. The participants are asked to evaluate its credibility, without imposing a particular method on them to do so. This means that they were free (and encouraged) to employ their own approach for this task. While doing this, the participants were asked to think aloud following standard think-aloud instructions (Ericsson & Simon, 1984). The participants were not allowed to navigate away from the article during the task. No time limit was set.

Design

A 3 (student group) × 2 (familiarity) × 2 (article quality) mixed design was applied for the experiment. Student group (high school, undergraduate, postgraduate) was a between-subjects factor, whereas familiarity (familiar/unfamiliar) and article quality (high/low) were both within-subject factors. Each participant evaluated 10 articles in total.

Familiarity was manipulated by selecting articles to be used in the experiment for each participant individually. This was done on the basis of a telephone interview, conducted a few days before the actual experiment. In this interview the participants were asked for their personal interests and disinterests. Half of the articles were selected to be on familiar topics; the other half were on unfamiliar topics. Each article was only used once throughout the whole experiment. Familiarity alternated between trials, starting with a familiar topic.

TABLE 2. Quality classes according to the Wikipedia Editorial Team Assessment.

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

Note. Detailed descriptions are available at http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment.

Article quality was manipulated following the classification of the Wikipedia Editorial Team ("Wikipedia:Version 1.0 Editorial Team," 2012). Manual assessments of the quality are available for most of the articles on Wikipedia, resulting in a categorization into seven classes (Table 2). However, A-class articles are largely underrepresented on Wikipedia, which makes it virtually impossible to find articles on specific topics in this class. Therefore, it was excluded from the experiment, leaving a total of six classes. Articles in the three highest classes were considered high quality (Featured articles, Good articles, and B-class articles); articles in the three lowest classes were considered low quality (C-class articles, Start articles, and Stub articles). Article quality was randomized between trials.

Unfortunately, no classification of quality is available on the Dutch Wikipedia (apart from the Dutch equivalent of the "Featured articles," but these are very few). Instead, we applied the clear criteria of the Wikipedia Editorial Team to articles from the Dutch Wikipedia ourselves to distinguish high and low quality. To ensure the validity of this manipulation, interrater reliability was calculated after double-rating the selected articles. The result was a Cohen's kappa of .89 (Landis & Koch, 1977), which indicates a near perfect agreement.

The articles used in the experiment were presented exactly as they appeared on Wikipedia, with the exception of the removal of cues specific to Wikipedia, indicating diminished credibility (e.g., [citation needed] indications) or high credibility (e.g., bronze stars in Featured articles). The

DOI: 10.1002/asi

removal of such indicators ensured that the participants could only utilize cues from the information itself in their credibility evaluations rather than cues only valid in the domain of Wikipedia.

Procedure

Upon arrival, participants were provided a brief explanation of the experiment and asked to sign an informed consent. As all participating high school students were younger than 18 years of age, we also asked their parents or legal guardians to sign an informed consent in advance.

After signing, the participants had to fill in a short questionnaire regarding standard demographic features and their familiarity and experience with Wikipedia (on 7-point Likert scales) along with their quotidian usage. They were also asked to provide a short explanation of what Wikipedia is and how it works.

Following this questionnaire the participants were instructed on the Wikipedia Screening Task and the course of the experiment. The participants practiced the Wikipedia Screening Task and the think-aloud task during two practice trials. The articles used in these trials were "Barcelona" and "Titanic" for the high school students, and "Flat earth" and "Ethnography" for the undergraduates and postgraduates. Task performance was considered sufficient for all participants after two practice trials.

When the participants finished a trial they indicated this to the experimenter verbally, who then handed them a questionnaire on which perceived credibility and familiarity were measured on 7-point Likert scales. This was repeated 10 times for each participant, resulting in a total duration of ~90 minutes.

Data Analyses

All sessions were audiorecorded and transcribed afterwards. In a protocol analysis, all utterances regarding credibility were marked and categorized. Each utterance was coded on the following aspects:

- The component of the article to which the utterance referred (Introduction, Text, Table of Contents, Images, References, and Other).
- The strategy applied by the participant (Semantic, Surface). Note that the Source strategy of the 3S-model (Lucassen & Schraagen, 2011) was not used here, as the source remained constant throughout the experiment.
- Which feature of the component was mentioned by the participant (e.g., number of references, quality of the pictures). Ad-hoc categories were created for the features mentioned.

The protocol of one student in each group was double-coded by two experimenters. Based on this overlap, the interrater reliability was calculated. A Cohen's kappa of .87 (Landis & Koch, 1977) indicated a near-perfect agreement.

During the protocol analysis, it became apparent that the participants differed greatly in number of utterances. In order to ensure that expressive participants did not have a larger influence on the outcome of each group than the others, the number of utterances of each participant (i) in each category (n) was corrected. This was done by multiplying each number by the correction factor derived in the following formula:

$$correction\ factor = \frac{number\ of\ remarks_n / n}{number\ of\ remarks_i}$$

After this correction, the number of utterances was averaged over each group to create a coding scheme for each group.

Only nonparametric tests were performed on all data gathered on Likert scales, as they are assumed to be measuring at the ordinal rather than the nominal level (Jamieson, 2004).

Results

Familiarity Manipulation Check

The questionnaires after each article indicated that the manipulation of familiarity was successful. On a 1 to 7 familiarity scale, familiar topics were rated higher ($M = 5.20$, $SD = 0.92$) than unfamiliar topics ($M = 1.94$, $SD = 0.87$), $Z = 5.48$, $p < .001$. A more detailed analysis showed that this was the case for all participating groups (high school students, undergraduates, and postgraduates).

Credibility Evaluation

Table 3 shows the number of remarks indicating the application of a semantic or surface strategy in the credibility evaluations of our participants. Typical examples of remarks categorized as a semantic strategy were "Yes, I know this is true, because the things I know about it are in line with the text," and "I know this already, because I traveled by airplane last year." Remarks such as "There are images everywhere, which seems trustworthy to me," and "Every claim is referenced, that's a good thing" were typical for the surface strategy.

Participants evaluating articles on familiar topics used more semantic cues than when evaluating unfamiliar topics, χ^2 (1, $N = 931$) = 24.40, $p < .001$. This was the case for all participant groups, high school students: χ^2 (1, $N = 661$) = 11.05, $p < .01$; undergraduates: χ^2 (1, $N = 1{,}122$) = 41.74, $p < .001$; postgraduates: χ^2 (1, $N = 1{,}010$) = 29.43, $p < .001$.

Moreover, the participant groups differed in their application of the semantic and surface strategy regardless of familiarity, χ^2 (2, $N = 931$) = 111.35, $p < .001$. This effect was caused by the high school students using fewer surface features than the other groups. No difference was found between undergraduates and postgraduates.

Table 4 shows the number of remarks concerning the various components of the articles. Participants in the various groups considered the different components of the article to varying degrees, χ^2 (12, $N = 2{,}793$) = 435.85,

DOI: 10.1002/asi

TABLE 3. (Corrected) Number of remarks indicating semantic or surface strategy application by the participants in all three groups.

	Familiar		Unfamiliar		All	
	Semantic	Surface	Semantic	Surface	Semantic	Surface
High school	215 (63.4%)	124 (36.6%)	163 (50.7%)	159 (49.3%)	378 (57.2%)	283 (42.8%)
Undergraduates	241 (41.6%)	339 (58.4%)	127 (23.4%)	415 (76.6%)	368 (32.8%)	754 (67.2%)
Postgraduates	257 (43.1%)	338 (56.9%)	110 (26.5%)	305 (73.5%)	367 (36.3%)	643 (63.7%)
Average	238 (47.1%)	267 (52.9%)	133 (31.2%)	293 (68.8%)	371 (39.8%)	560 (60.2%)

Note. Percentages are given in parentheses.

TABLE 4. (Corrected) Number of remarks indicating the utilization of several components of the information by the participants in all three groups.

	Introduction	Text	Table of contents	Images	Internal links	References	Other
High school	18 (2.7%)	532 (80.5%)	3 (0.4%)	72 (10.9%)	11 (1.7%)	4 (0.5%)	21 (3.2%)
Undergraduates	49 (4.3%)	485 (43.2%)	38 (3.4%)	140 (12.5%)	33 (2.9%)	319 (28.4%)	59 (5.2%)
Postgraduates	99 (9.8%)	408 (40.4%)	38 (3.7%)	66 (6.5%)	27 (2.7%)	337 (33.4%)	34 (3.4%)
Average	166 (5.9%)	1425 (51.0%)	79 (2.8%)	278 (10.0%)	71 (2.5%)	660 (23.6%)	114 (4.1%)

Note. Percentages are given in parentheses.

TABLE 5. Key features used by participants in each group.

	High school	Undergraduates	Postgraduates
Factual accuracy	X	X	X
Completeness	X	X	X
Images	X	X	X
Length of text	X	X	X
Writing style	X	X	X
Quality of text	X	X	X
Scope of text	X	X	X
Understandability	X	X	X
References		X	X
Objectivity		X	X
Structure		X	X
Statistics		X	

Note. Strategies were included if at least 50% of the participants in that group applied the strategy at least once.

$p < .001$. Post-hoc analyses showed that this was caused by postgraduates having fewer remarks on images and more on the introduction than the other groups. The number of remarks on references differed between all groups, increasing with education level. Finally, high school students mentioned the component "text" more and "table of contents" less than the other groups.

Table 5 shows the key features used by each group. As can be seen in Table 4, high school students had a smaller arsenal of strategies to evaluate credibility than the other groups. Some evident strategies are mentioned by all users (e.g., factual accuracy), but other strategies such as considering the references and the objectivity of the information were only mentioned by undergraduates and postgraduates.

Trust

Table 6 shows trust in the information of all participants in all conditions. No effect of student group on trust was found, χ^2 (2, $N = 40$) = .21, $p = .90$. This indicates that high school students, undergraduates, and postgraduates all have similar trust in Wikipedia.

Moreover, no effect of familiarity on trust was found, $Z = 1.68$, $p = .09$. This was also the case for each individual student group (high school: $Z = 1.09$, $p = .28$; undergraduates: $Z = 1.30$, $p = .19$; postgraduates: $Z = 0.66$, $p = .51$).

Quality had a significant effect on trust: high-quality articles were trusted more than low-quality articles ($Z = 3.62$, $p < .01$). However, a more detailed analysis showed that this was only the case for undergraduates ($Z = 2.67$, $p < .01$) and postgraduates ($Z = 2.84$, $p < .01$), but not for high school students ($Z = 1.37$, $p = .17$).

Discussion

In this study, the influence of domain expertise and information skills on credibility evaluation and trust was

DOI: 10.1002/asi

TABLE 6. Trust in the information on 7-point Likert scales in all conditions.

	Familiar			Unfamiliar			All		
	HQ	LQ	All	HQ	LQ	All	HQ	LQ	All
High school	5.77	5.00	5.29	5.62	4.44	5.03	5.69	4.74	5.23
	(1.03)	(1.78)	(1.09)	(0.97)	(1.96)	(0.96)	(0.71)	(1.82)	(0.94)
Undergraduates	5.86	4.63	5.28	5.71	4.38	5.00	5.76	4.52	5.14
	(0.72)	(1.31)	(0.75)	(0.56)	(1.05)	(0.72)	(0.52)	(1.05)	(0.65)
Postgraduates	5.46	4.78	5.16	5.33	4.82	5.05	5.37	4.84	5.11
	(0.49)	(1.05)	(0.63)	(0.61)	(0.89)	(0.69)	(0.44)	(0.76)	(0.54)
Average	5.69	4.80	5.24	5.55	4.55	5.03	5.61	4.70	5.16
	(0.75)	(1.38)	(0.82)	(0.71)	(1.30)	(0.79)	(0.56)	(1.21)	(0.71)

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

examined. The results supported the updated 3S-model. It was found that users with domain expertise tended to focus more on semantic features than users without domain expertise. Moreover, surface features were used more by users with better information skills. Information quality was manipulated following the classification of the Wikipedia Editorial Team. We hypothesized that this would be mainly visible at the surface level of the articles. Our experiment confirmed that indeed only trust of users with better information skills was influenced by the quality manipulation (i.e., only undergraduates and postgraduates, not high school students). As expected, domain expertise had no influence on trust in high- or low-quality articles, as low-quality articles are also expected to be free of (large) factual errors.

The main contribution of this study is that the enabling character of domain expertise and information skills has been demonstrated, together with the influence of the corresponding information features on trust. In the original study on the 3S-model (Lucassen & Schraagen, 2011), it was shown that trust of domain experts was influenced when a semantic feature (enabled for domain experts) indicated low credibility. Now, we also demonstrate that when surface features indicate lower credibility, this only has an influence on users who have sufficient information skills. Hence, only undergraduates and postgraduates were influenced, whereas high school students were not.

This observation is very much in line with prominence-interpretation theory (Fogg, 2003), which states that each cue in a piece of information has a certain prominence for a given user. Only when a cue is prominent can it be interpreted by the user (i.e., have consequences for credibility), and influence trust. The key addition of the 3S-model in comparison to prominence-interpretation theory is that we attribute specific user characteristics to specific information features.

In this experiment, we showed that people with knowledge of the topic evaluate the credibility of information differently than people without such knowledge. The key difference is the utilization of semantic features, such as the accuracy of information. Novices are not able to compare presented information with their preexisting knowledge, which leads them to the consideration of other, surface features. Interestingly, one does not have to be an absolute domain expert to apply the "semantic strategy" of credibility evaluation. Whereas in Lucassen and Schraagen (2011), domain experts were self-selected from various Internet forums on automotive technology, in this experiment familiarity was merely manipulated by asking the participants for their topics of interest. This does not ensure a substantial level of expertise at all. Still, the influence of familiarity at this level on credibility evaluation was made quite clear. Participants familiar with the topic at hand used nearly twice as many semantic features in their credibility evaluations as participants unfamiliar with the topic (except for high school students, to be discussed later).

Interestingly, when users encounter information on a familiar topic, they do not shift to semantic features completely. Instead, they apply a combination of surface and semantic strategies to evaluate credibility. This means that familiar users (with sufficient information skills) are best equipped to evaluate credibility in a meaningful manner. However, this experiment merely indicated the capabilities of various users to evaluate, which may differ from their actual behavior in real life. As predicted by Metzger (2007), the motivation of users primarily determines to what extent credibility is evaluated. This experiment showed what they are capable of when they are motivated.

As stated earlier, the shift towards semantic features when evaluating familiar information was much less distinctive for high school students than for the other two groups. We do not attribute this to an unexpected high level of expertise in unfamiliar topics (the familiarity manipulation proved successful), but to a low level of information skills. We argued before that the most salient strategy for credibility evaluation is to consider the factual accuracy. This is also what the high school students did. However, when evaluating information on unfamiliar topics, this strategy is quite unsuccessful. We also observed this in many participants remarking that they felt unable to evaluate the article at hand, as they did not know anything about the topic (e.g., "If you don't know anything about it, it is tempting to believe the information is correct." or "It doesn't ring a bell, it could

DOI: 10.1002/asi

be true."). Participants with better information skills worked around this deficit by considering various alternative (surface) features. However, high school students were not able to do so, due to their limited information skills, which meant that a large portion (about half) of the remarks still included semantic features (albeit unsuccessful). This was reflected in the trust of high school students in the information; as opposed to the other groups, no difference in trust was observed between high-quality and low-quality information. The key surface feature that high school students did not consider at all as opposed to the other groups was references. High school students were not at all aware of the importance of references, whereas a large part of the remarks of undergraduates and postgraduates considered this feature (about 30%). This replicates the finding of Lucassen et al. (2011), who found that undergraduates consider the references of information on various levels.

The limited information skills of high school students could lead one to believe that they could still perform a meaningful credibility evaluation on familiar topics, as they can bring their knowledge of the topic at hand to bear. However, no influence of information quality on trust was found for high school students, regardless of their familiarity. It could be argued that this is due to their limited domain expertise, also with familiar topics. However, a more plausible explanation can be found in the nature of the quality manipulation at hand. We decided to replicate normal quality fluctuations that can be observed on Wikipedia. However, these fluctuations can primarily be found at the surface level, as the information is expected to be generally factually accurate (Giles, 2005) also in low-quality articles. High school students utilize a lot less surface features in their evaluations, which means they did not notice the differences in quality.

The manner of manipulating information quality also explains why, despite earlier findings (Chesney, 2006; Eastin, 2001; Self, 1996), familiarity had no influence on trust. It can be expected that most of the articles used in this experiment were factually accurate. Thus, no negative influence of knowledge on the information is expected. However, given the overall high trust in the presented information (>5 on a 1 to 7 scale), it is questionable whether familiarity would increase trust even more. A study in which the role of familiarity is examined in trust in information of more questionable credibility would be of interest to further explore this topic.

No effect of participant group on trust was found. This means that trust in Wikipedia is the same for high school students, undergraduates, and postgraduates. This is remarkable because knowledge of the open-editing system behind Wikipedia (largely absent in high school students) could lead to less trust. On the other hand, accumulating positive experiences with Wikipedia may increase trust in this source (Lucassen & Schraagen, 2011). This would indicate that the strategy of considering the source of information was also applied, but implicitly, as no participants mentioned this in the think-aloud protocols (Taraborelli, 2008).

Limitations

A few limitations should be kept in mind regarding the interpretation of the results of this study. The three participating groups of students were selected on the basis of their expected level of information skills. We have shown that this had a direct influence on credibility evaluation. However, other factors will also inevitably vary among these groups (e.g., age). These factors may act as confounding variables. However, we would argue that an isolated approach to varying information skills (e.g., training half of a coherent group of participants with low information skills) does not add to the external validity of the study. We also suggest that in future studies concerning differences in information skills among groups, information skills could be measured to have a better grasp on such differences (e.g., project SAILS; Rumble & Noe, 2009).

In this study, Wikipedia served as an information source for our stimuli. This online encyclopedia is always a great case study, as information quality is generally very high (Giles, 2005), but changeable (e.g., Cross, 2006; Dooley, 2010). However, certain characteristics of this source may limit the potential for generalization to other sources (online and offline). An example of such a characteristic is the open editing model behind Wikipedia; this mechanism may cause users to approach the information differently (e.g., in a more skeptical manner). For future research, it is important to verify the validity of the proposed 3S-model in different contexts, such as other websites, or offline sources (e.g., books, newspapers).

The think-aloud method is a great tool to gain insight into the task performance of participants. It should be noted, however, that in this study the participants were explicitly asked to evaluate credibility, whereas normally this is a subset of a larger task set (i.e., finding and evaluating information). Therefore, the observed behavior in this experiment should not be interpreted as the way users always perform credibility evaluation. The degree to which credibility is actually being evaluated may vary strongly (Metzger, 2007). The behavior we observed in this study can rather be seen as credibility evaluation under optimal circumstances (in terms of motivation and ability). In real life, users may pick a few strategies from the set we found, depending on the context of the information.

Further Research

This study has shed more light on the role of user characteristics in online credibility evaluation. Additional validation was found for the 3S-model (Lucassen & Schraagen, 2011). However, this study primarily aimed at validating the semantic and surface components of the model. Future studies should also focus on the third strategy, considering the source of information.

Acknowledgments

The authors thank Andreas Bremer, Knut Jägersberg, and Koen Remmerswaal for their efforts in gathering the data.

DOI: 10.1002/asi

Also, we thank C.S.G. Dingstede for allowing us to perform the experiments with high school students at their school.

References

Alexander, J.E., & Tate, M.A. (1999). Web wisdom; how to evaluate and create information quality on the web, 1st ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. Computers in Human Behavior, 21, 487–508.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology, 39, 752–766.

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

Cross, T. (2006). Puppy smoothies: Improving the reliability of open, collaborative wikis. First Monday, 11.

Dooley, P.L. (2010). Wikipedia and the two-faced professoriate. In Proceedings of the 6th International Symposium on Wikis and Open Collaboration, WikiSym '10 (pp. 1–2). New York: ACM.

Eastin, M.S. (2001). Credibility assessments of online health information: The effects of source expertise and knowledge of content. Journal of Computer-Mediated Communication, 6.

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

Flanagin, A.J., & Metzger, M.J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. New Media & Society, 9, 319–342.

Fogg, B.J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. CHI '03 extended abstracts on human factors in computing systems, CHI EA '03 (pp. 722–723). New York: ACM.

Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J., & Tauber, E.R. (2003). How do users evaluate the credibility of web sites? A study with over 2,500 participants. In Proceedings of the 2003 Conference on Designing for User Experiences, DUX '03 (pp. 1–15). New York: ACM.

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

Giles, J. (2005). Internet encyclopaedias go head to head. Nature, 438, 900–901.

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

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education, 38, 1217–1218.

Julien, H., & Barker, S. (2009). How high-school students find and evaluate scientific information: A basis for information literacy skills development. Library & Information Science Research, 31, 12–17.

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.

Klein, G.A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. Proceedings of the Human Factors Society (pp. 576–580). Human Factors and Ergonomics Society.

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

Lucassen, T., Noordzij, M.L., & Schraagen, J.M. (2011). Reference blindness: The influence of references on trust in Wikipedia. In ACM WebSci '11. New York: ACM.

Lucassen, T., & Schraagen, J.M. (2010). Trust in Wikipedia: How users trust information from an unknown source. Proceedings of the 4th Workshop on Information Credibility, WICOW '10 (pp. 19–26). New York: 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, 1232–1242.

Magnus, P.D. (2009). On trusting Wikipedia. Episteme, 6, 74–90.

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734.

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.

Rajagopalan, M.S., Khanna, V., Stott, M., Leiter, Y., Showalter, T.N., Dicker, A., & Lawrence, Y.R. (2010). Accuracy of cancer information on the internet: A comparison of a wiki with a professionally maintained database. Journal of Clinical Oncology (Meeting Abstracts), 28, 6058.

Robins, D., & Holmes, J. (2008). Aesthetics and credibility in web site design. Information Processing & Management, 44, 386–399.

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.

Rumble, J., & Noe, N. (2009). Project SAILS: Launching information literacy assessment across university waters. Technical Services Quarterly, 26, 287–298.

Self, C.S. (1996). Credibility. In M. Salwen & D. Stacks (Eds.), An integrated approach to communication theory and research. Mahwah, NJ: Lawrence Erlbaum Associates.

Taraborelli, D. (2008). How the web is changing the way we trust. In K. Waelbers, A. Briggle, & P. Breg (Eds.), Current issues in computing and philosophy (pp. 194–204). Amsterdam: IOS Press.

Viégas, F., Wattenberg, M., & Kushal, D. (2004). Studying cooperation and conflict between authors with history flow visualization. Proceedings of the 2004 Conference on Human Factors in Computing Systems. New York: ACM.

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. (2009). How students evaluate information and sources when searching the world wide web for information. Computers & Education, 52, 234–246.

Wikipedia:Version 1.0 Editorial Team. (n.d.). In Wikipedia. Retrieved from http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team

Wilkinson, D.M., & Huberman, B.A. (2007). Cooperation and quality in Wikipedia. In Proceedings of the 2007 International Symposium on Wikis, WikiSym '07 (pp. 157–164). New York: ACM.

Yaari, E., Baruchson-Arbib, S., & Bar-Ilan, J. (2011). Information quality assessment of community generated content: A user study of Wikipedia. Journal of Information Science, 37, 487–498.

Appendix

The original 3S-model as proposed by Lucassen and Schraagen (2011) (Figure 2).

DOI: 10.1002/asi