Teun Lucassen

Trust in Online Information

Teun Lucassen

Doctoral committee:

Chair:	Prof. dr. K. I. van Oudenhoven-van der Zee
Promotor:	Prof. dr. J. M. C. Schraagen
Members:	Prof. dr. D. K. J. Heylen
	Dr. P-.P. van Maanen
	Prof. dr. C. J. H. Midden
	Prof. dr. M. A. Neerincx
	Prof. dr. ing. W. B. Verwey
	Dr. A. Walraven

TRUST IN ONLINE INFORMATION

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 1 maart 2013 om 16.45 uur

door

Teun Lucassen
geboren op 12 april 1983
te Meppel

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. J. M. C. Schraagen

CTIT Ph.D. Thesis Series No. 13-242
Centre for Telematics and Information Technology
P.O. Box 217, 7500 AE
Enschede, The Netherlands

ISBN: 978-90-365-3485-7
ISSN: 1381-3617 (CTIT Ph. D. Thesis Series No. 13-242)
DOI: 10.3990/1.9789036534857

Cover: Teun Lucassen
Print: Ipskamp, Enschede

Table of Contents

Chapter 1

General Introduction

1. Introduction

The Internet has enriched our lives. Try to imagine how we did certain jobs before we had the ability to go online: How did we for example book a flight to our holiday destination? The answer seems straightforward: we went to a local travel agency, and let them book one. But how did we know that we had booked a flight in accordance with all our preferences (leg space!), and for a competitive price? We could have visited a few more agencies in our home town, but still, a cheap agency a few miles away could have had an even better offer. Quite frankly, we didn't know.

A different example: How did we ever buy a used car in the eighties or early nineties? I for one am very specific about my preferred make, model, and build when it comes to cars. The market for second hand cars is large, but how could we trace the right one for us without having the Internet? Basically, we were dependent on a few local dealerships, or perhaps the classifieds in newspapers. Still, this gave you the choice between a few dozens rather than the thousands and thousands of options that we have nowadays. These are only two random examples, but just think about what the Internet has done for basically every aspect of our lives, such as education, healthcare, dating, shopping, entertainment, research, and so on.

Of course, with the many upsides that the world of online information has provided us, downsides are bound to come along with them. The Internet is for example not especially well-organized. The right information is out there somewhere, but to actually find it may not always be an easy task. A completely new skill set to search and find information is required, and not everyone has been able to adapt (consider the digital divide, Palfrey & Gasser, 2008). Moreover, once we have found the information we looked for, how do we know that we can trust it?

The second wave of Internet technology (Web 2.0; Cormode & Krishnamurthy, 2008) has made matters worse by enabling and encouraging end-user contributions (i.e., user-generated content; Alexander, 2006). That means that in many cases, anyone could have put the information online. But how do we know that this person has the right motives? Perhaps this person does not want you to have the correct information; consider for instance corrupt regimes in foreign countries, or shady Internet vendors, who just want to make a quick sale. Their motives may not be beneficial to us, but how do we identify such parties? And moreover, how do we know that the person who put the information online

has sufficient knowledge about the topic? We often do not know whether he or she is a high school student, or an established professor in the domain of the information.

What this boils down to is that Internet users somehow need to evaluate the credibility of information they find online. However, two problems can be identified immediately. The first is that compared to the pre-Internet era, we have far fewer professional gatekeepers (e.g., editors, journalists) checking the quality of the information for us (Flanagin & Metzger, 2007). This means that the task of evaluating credibility has shifted towards the end-users, who are often not trained for this. Second, the traditional strategy of considering the source of information (Chaiken & Maheswaran, 1994; Sundar, 2008) has become problematic, as the actual source is often unknown in online environments (Lim & Kwon, 2010). Moreover, sources are often "layered", meaning that information travels through multiple sources before reaching the end-user (Sundar & Nass, 2001; Kang, Bae, Zhang, & Sundar, 2011). This makes it hard to identify which source is actually responsible for the credibility of the information.

The problems in evaluating credibility that Internet technology has brought along have resulted in a lot of research in various disciplines (Flanagin & Metzger, 2007). Various aspects of trust in online information have been investigated, including informational features (Kelton, Fleischman, & Wallace, 2008), contextual features (Metzger, 2007), and user characteristics (Fogg, 2003; Sundar, 2008). In this dissertation, I aim at contributing to the existing knowledge on the influence of various user characteristics on credibility evaluation. In the remainder of this chapter, I start by providing a definition of what we exactly mean by trust and credibility in an online context, followed by a brief discussion of current attempts to explain trust in online information and the methodology applied in these attempts. In Chapters 2 to 5, I focus on three particular user characteristics, namely domain expertise, information skills, and source experience, which I integrate into one model of trust, coined the 3S-model. In Chapter 6, I extend these characteristics by a general propensity to trust, and the influence of trust in a medium on trust in information. In Chapters 7 and 8, I explore a potential solution to the problems that users have with evaluating credibility by offering them advice through a decision support system. I conclude this dissertation with a general discussion of the findings, along with implications for research and practice , and suggestions for future research.

1.1 Trust and Credibility

Before addressing the various aspects of trust in online environments, it is important to establish a working definition of the concept of trust itself. A general definition of trust, irrespective of the environment it is applied in, is given by Mayer, Davis, and Schoorman (1995):

> *The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (p. 712)*

From this definition we can derive that trust is fundamentally a relational construct applicable to two or more people. Trust in information supplied by another party can thus be interpreted as trust in the other party that the supplied information is correct. Kelton et al. (2008) distinguish four levels at which trust can be studied, namely a) *individual* (as a personality trait), b) *interpersonal* (one person trusting another), c) *relational* (property in a mutual relationship), and d) *societal* (characteristic of a society). In studies on trust in information, the focus is largely on interpersonal trust. A higher level of trust, such as relational trust is normally not necessary in information exchanges, as there is normally no need for the author to trust the user of information (although exceptions can be thought of, consider for instance the disclosure of private information).

Corritore, Kracher, and Wiedenbeck (2003) have provided a definition of trust that is specifically adapted to online environments. In this definition, the "other party" is taken out of the equation, as this is often unknown or unclear (Lim & Kwon, 2010). It thus focuses only on the expectation of the Internet user:

> *An attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited. (p. 740)*

An aspect of trust that we see in both definitions of Mayer et al. (1995) and Corritore et al. (2003) is *vulnerability*. Internet users may be vulnerable when they interact with others. This implies that a certain risk is taken when someone or something is trusted. In some situations, users may feel the need to reduce this risk, as the consequences of poor information may be large (Metzger, 2007). Consider for instance information used in decisions on personal medical care; the consequences of poor information may be quite severe. In such situations, the user may attempt to find cues about the *credibility* of

information, thereby reducing the risk that is taken.

Credibility may be defined as the believability of information (Fogg & Tseng, 1999). Trust and credibility are often confused or used interchangeably in the literature (Fogg & Tseng, 1999), but we treat trust as an act of a user, whereas credibility is a property of information. In cognitive psychology, credibility is determined by two main elements, namely trustworthiness and expertise (Hovland & Weis, 1951, Hovland, Janis, & Kelley, 1981; Self, 2009). The first refers to the willingness to provide correct information (intention), whereas the second refers to the ability to provide correct information (knowledge). The task of credibility evaluation is also described regularly as estimating information quality (Fogg & Tseng, 1999; Eastin, 2001; Metzger, 2007; Sundar, 2008).

1.2 Trust in Online Environments

In online environments, the concepts of trust and credibility are especially salient in comparison to traditional (offline) information sources for several reasons. To start with, information on the Internet is prone to alteration (Johnson & Kaye, 1998; Alexander & Tate, 1999; Flanagin & Metzger, 2000). Flanagin and Metzger (2007) posited that online information is probably less reliable due to structural and editorial features of the Internet. As stated earlier, there is a lack of professional gatekeepers who monitor the quality of the information. Moreover, various information genres tend to converge on the Internet. It has, for instance, become harder and harder to distinguish advertisements from other information. Also, because the Internet is relatively young, most sources lack a reliable reputation. Kelton et al. (2008) note that a lot of environmental and behavioral cues are absent in online environments.

According to Sundar (2008), the traditional heuristic to determine the credibility of information is to look at the source. However, this approach is "murky" in online contexts. Multiple authors may contribute to one piece of information (consider for instance Wikipedia), which makes it impossible to hold one individual responsible for the information. In some cases, the source may be unknown or anonymous. Also, sources are often layered (Sundar & Nass, 2001), which means that information travels through multiple sources before reaching the end-user. When someone writes a blog on an item on a news website, which is in turn largely based on an article on Wikipedia, who can be kept responsible for the credibility of the information? Research has shown that users with low involvement mainly consider the proximate source (Kang, Bae, Zhang, & Sundar, 2011).

A final complicating factor in credibility evaluation of online information is the sheer number of alternative sources. It is very hard to select the credible ones (Hilligoss & Rieh, 2008). Following the theory of bounded rationality (Simon, 1982/1997), users are prone to follow a "satisficing" approach (Simon, 1956) to avoid spending a lot of time and effort in this process, which may have consequences for the quality of their judgments.

In conclusion, credibility evaluation is a particularly difficult task in online environments and therefore worthy of further research attention.

1.3 Models of Online Trust

Several attempts have been made to explain how trust in online information is formed. Next, I discuss a few of them, and identify a knowledge gap that I intend to address.

Corritore et al. (2003) argue that two main categories of factors impact trust, namely external factors (implicit and explicit) and a user's perception of these factors. The key factors in perception are credibility and risk, both influenced by ease of use. A combination of the perception of credibility and risk leads to trust in a particular piece of information.

The distinction between external factors and their perception is extended by Fogg (2003) in his Prominence-Interpretation theory. Following this theory, information can be divided into various elements, which may all have an impact on a user's trust. For an element to have impact, it first needs to be noticed by a user (prominence). At least five factors influence the likelihood that an element is noticed, namely, involvement, topic, task, experience, and individual differences. Once an element is prominent for a user, it needs to be interpreted (e.g., "a lot a references is good for credibility"). According to Fogg (2003), this is affected by assumptions and skill or knowledge of a user, as well as by contextual factors. Adding up all prominent and interpreted elements of a piece of information leads to a credibility evaluation.

Of course, the number of elements incorporated in a credibility evaluation may vary. Dual processing theory (Chaiken, 1980; Chaiken & Maheswaran, 1994) may shed more light on the effort that is put into evaluating credibility. Metzger (2007) has proposed a dual processing model of Web site credibility assessment. In this model, two key factors determine whether a heuristic (peripheral) or systematic (central) evaluation is performed. When being confronted with a Web site, users need to have a motivation to evaluate the information. According to Metzger (2007), this motivation "stems from the

consequentiality of receiving low-quality, unreliable, or inaccurate information" (p. 2087). Unmotivated users will not perform a credibility evaluation at all, or at most a heuristic evaluation. For motivated users, the employment of a heuristic or systematic evaluation depends on the ability of the user. Ability refers to "the users' knowledge about how to evaluate online information" (p. 2087).

Despite the notion of systematic processing in the model of Metzger (2007), it can be argued that credibility evaluation is always heuristic to a certain extent. As can be inferred from the definitions given earlier (Mayer et al., 1995; Corritore et al., 2003), trust involves taking a certain risk. Full systematic processing would mean that every aspect relevant for credibility is evaluated, reducing this risk to zero. This means that trust is replaced by certainty. Hence, credibility evaluation is always heuristic to a certain extent.

The notion of the heuristic character of credibility evaluation is supported by the MAIN model, proposed by Sundar (2008). In this model, it is posited that the online environment has yielded certain affordances that can convey credibility cues. These affordances can be categorized in Modality, Agency, Interactivity, and Navigability (hence, MAIN). Each of the affordances gives access to a list of heuristics which in turn are predictive of the quality of the information. Similar to Fogg (2003), combining a number of these heuristics results in a credibility judgment.

Heuristics also play a large role in the unifying framework of credibility assessment by Hilligoss and Rieh (2008). Three levels between the information seeker (user) and information object are defined, namely the construct level, the heuristics level, and the interaction level. The construct level involves the personal definition of credibility of a user. This means that this model acknowledges that the way a credibility evaluation is performed varies between different users, as they utilize different definitions. The second level represents the general heuristics a user applies, even before considering the information itself (e.g., "information from the Internet is never credible"). The third and final level concerns the actual interaction between the information and the user. Interaction can be divided into the utilization of content cues, source cues, and peripheral information object cues. Systematic evaluation can occur only at this level.

The unifying framework by Hilligoss and Rieh (2008) reveals an essential topic also visible in other models of online trust, namely the influence of user characteristics on credibility evaluation. Hilligoss and Rieh (2008) argue that definitions of credibility vary between users. Metzger (2007) talks about the motivation and ability of a user, which may also

vary. Fogg (2003) argues that prominence and interpretation of credibility cues will differ between persons. Also, Corritore et al. (2003) theorize about perception of external factors, which is bound to introduce differences among users. However, while arguments have been made about what these differences are (e.g., need for cognition, learning style, and literacy level; Fogg, 2003), empirical evidence for their effect on the process of credibility evaluation and ultimately trust is largely absent. I intend to fill this gap in this dissertation, predominantly led by the three elements of the interaction level in the framework by Hilligoss and Rieh (2008): the use of content cues, source cues, and peripheral information object cues. This leads to the central research question of this dissertation:

> *How do user characteristics influence credibility evaluation and trust in online information?*

In this dissertation, I will focus on three user characteristics that I hypothesize to be directly influential on the three types of interaction cues proposed by Hilligoss and Rieh (2008). These are domain expertise on the topic at hand, information skills, and source experience (Chapter 2 to 5). In Chapter 6, I expand these characteristics with a more general propensity to trust (McKnight, Kacmar, & Choudhury, 2004) and trust in a medium (Sundar & Nass, 2001).

1.4 Supporting Credibility Evaluation

The existing literature on trust in online information not only explains how users evaluate credibility, it also demonstrates that in many situations, users have difficulties with this task (e.g., Walraven, Brand-Gruwel, & Boshuizen, 2009). This may for instance be caused by a lack of motivation or ability to evaluate (Metzger, 2007).

One potential method to improve credibility evaluations is to provide decision support (e.g., Adler et al., 2008). Such support could give advice about the credibility of the information, aiding the user in his or her decision making process (Lee & See, 2004). Several attempts have already been made in the domain of Wikipedia to aid users in evaluating the articles of this encyclopedia. WikiTrust, for instance, colors the background of each word according to its credibility, based on how many edits to the article that word has survived (Adler, et al., 2008). Chevalier, Huot, and Fekete (2010) offer several indicators to the user, such as the length of the text, number of links, and the number of contributors to the article. Suh, Chi, Kittur, and Pendleton (2008) created a visualization that gives an insight in the edit history of each Wikipedia article, and Korsgaard and Jensen (2009) have

proposed a reputation system in which users can rate the articles themselves.

A brief analysis of these attempts shows that they are primarily technology-driven; the support is based on features that can be measured automatically. However, studies on credibility evaluation behavior have shown that not all of the features considered by Internet users can be measured so easily (Yaari, Baruchson-Arbib, & Bar-Ilan, 2011). This means that it is likely that there is a discrepancy between the features employed by decision support systems and the mental model of the user, which may be disadvantageous for user acceptance. This has, for instance, been shown by Lucassen and Schraagen (2011a); participants did not perceive added value of WikiTrust, because they had difficulties in incorporating the advice into their own evaluation.

In the second part of this dissertation (Chapter 7 and 8), I will investigate the application of decision support in credibility evaluation from a user's perspective. Rather than searching for new measurable cues that correlate well with credibility (or information quality), I will attempt to find out the prerequisites of decision support to be accepted by the user. To do so, multiple systems representing the various implementation options will give simulated advice to the user.

2. Methodology

Several approaches have been taken to study trust in online environments. Analysis of 30 studies related to this topic has revealed some patterns in the methodology applied. Table 1 shows a summary of the relevant studies. Based on this table, I will now discuss the choices made in this dissertation in method, participants, and task setting.

2.1 Method

As can be derived from Table 1, the use of questionnaires in credibility evaluation studies is widely accepted. Two-thirds of the studies applied this method in an online (n = 14) or offline (n = 5) context. In 15 of these studies, certain stimuli were administered to the participants, with questions regarding these stimuli afterwards (e.g., "How credible do you find this website?", Robins & Holmes, 2008). In the remaining five studies, a questionnaire without stimuli was used, which means that the questions are of a more general nature (e.g., "What do you think is the credibility of the Internet?", Dutton & Shepherd, 2006).

Table 1: Summary of the research methodology of 30 studies on trust in online information.

Study	Method	Participants	Task setting
Brand-Gruwel, Wopereis, & Vermetten (2005)	Offline questionnaire with stimuli, Think-aloud	Students	Information problem solving
Chesney (2006)	Expert evaluation	Academics	Credibility evaluation
Chevalier, Huot, & Fekete (2010)	Offline questionnaire with stimuli	General population, random	Credibility evaluation
Dutton & Shepherd (2006)	Online questionnaire without stimuli	General population, representative	-
Flanagin & Metzger (2007)	Online questionnaire with stimuli	General population, random	Natural browsing
Flanagin & Metzger (2011)	Online questionnaire with stimuli	General population, representative	Credibility evaluation
Fogg et al. (2003)	Online questionnaire with stimuli	General population, random	Credibility evaluation
Giles (2005)	Expert evaluation	Academics	Credibility evaluation
Hargittai, Fullerton, Menchen-Trevino, & Thomas (2010)	Online questionnaire without stimuli, Naturalistic observation, Interview	Students	Information problem solving
Head & Eisenberg (2010)	Online questionnaire without stimuli, Focus groups	Students	-
Hilligoss & Rieh (2008)	Naturalistic observation, Interview	Students	Natural browsing
Julien & Barker (2009)	Interview	Students	Information problem solving
Kang, Bae, Zhang, & Sundar (2011)	Online questionnaire with stimuli	Students	Credibility evaluation
Kim & Sundar (2011)	Offline questionnaire with stimuli	Students	Credibility evaluation
Kittur, Suh, & Chi (2008)	Online questionnaire with stimuli	General population, random	Credibility evaluation
Kubiszewski, Noordewier, & Costanza (2011)	Online questionnaire with stimuli	Students; Academics	Credibility evaluation
Lim & Kwon (2010), Lim (2009)	Online questionnaire without stimuli	Students	-
Liu (2004)	Offline questionnaire without stimuli	Students	-
McKnight, Kacmar, & Choudhury (2004)	Online questionnaire with stimuli	Students	Credibility evaluation
McKnight & Kacmar (2006)	Online questionnaire with stimuli	Students	Natural browsing
Metzger, Flanagin, & Medders (2010)	Focus groups	General population, random	-

Rajagopalan, Khanna, Stott, Leiter, Showalter, Dicker, & Lawrence (2010)	Expert evaluation	Academics	Credibility evaluation
Robins & Holmes (2008)	Offline questionnaire with stimuli	Students	Credibility evaluation
Robins, Holmes, & Stansbury (2010)	Offline questionnaire with stimuli	General population, random	Credibility evaluation
Sundar & Nass (2001)	Offline questionnaire with stimuli	Students	Credibility evaluation
Sundar, Knobloch-Westerwick, & Hastall (2007)	Online questionnaire with stimuli	Students	Credibility evaluation
Walraven, Brand-Gruwel, & Boshuizen (2009)	Think-aloud	Students	Information problem solving
White, Pahl, Buehner, & Haye (2003)	Online questionnaire with stimuli	Students; General population, random	Information problem solving
Yaari, Baruchson-Arbib, & Bar-Ilan (2011)	Think-aloud	Students	Credibility evaluation

Other, less popular methods are think-aloud (Ericsson & Simon, 1984), expert evaluation, interviews (often in combination with other methods), focus groups, and naturalistic observation.

As stated earlier, my research goal is to investigate the relationship between user characteristics and credibility evaluation. In order to demonstrate the direct effect of certain characteristics on the way credibility evaluation is performed, one needs to to 1) manipulate or control these user characteristics and 2) manipulate or control the information features that are expected to be related to the user characteristics. Therefore, the application of questionnaires without the administration of stimuli is not a suitable approach. Neither are naturalistic observation, interviews, and focus groups suitable approaches for my purposes, as these also do not allow manipulation or control over the stimulus material.

From the methods listed above, this leaves the application of questionnaires with stimuli, and the think-aloud method. A major advantage of questionnaires is that it is relatively easy to include a high number of participants because they can often be recruited online. Also, participants in specific target groups can easily be found on corresponding forums (e.g., car enthusiasts, gamers). Use of a large number of representative users results in higher external validity of the results obtained. The downside of this method is that the accountability of the participants is very low, that is, the experimenter has little or no control over the variables involved during task performance. Think-aloud on the other hand, results in more control during task performance and very detailed information

about the thought processes of the participants (Ericsson & Simon, 1984), but is generally not suitable for use with hundreds of participants due to its time-consuming character.

Rather than choosing between one or the other of these two methods suitable for the administration of (manipulated) stimuli, I will apply both methods in the experiments discussed in this dissertation. Online questionnaires will be used when a user group with specific expertise is needed (Chapter 2) or when a broad spectrum of Internet users is helpful (Chapter 6), whereas think-aloud will be applied when more detailed information about the cognitive processes of the participants is required (Chapter 3). In Chapters 4, 5, 7, and 8, questionnaires will be deployed, but in a controlled lab environment. This improves the accountability of the participants. This multi-method approach should combine the best of both worlds: external validity as well as control of all sources of variance.

Questionnaires (predominantly online) are also the *modus operandi* in trust research in the domain of e-commerce (Grabner-Kräuter & Kaluscha, 2003). However, in their review on empirical studies on this topic, one major concern was expressed, namely that trust is measured very differently across various studies, without a theoretical underpinning or rigorous validation. Nevertheless, since I am especially interested in studying the differences in credibility evaluation between users, I will not use validated questionnaires with multiple scales to measure trust, as this may lead participants towards the employment of certain information features they would not consider themselves. Consider for instance the question "How much do you trust the institutes and people 'running the Internet'?", derived from Dutton and Shepherd (2006). This is a valid question to measure trust in the Internet, but it may lead participants to consider the credibility of the people who put information online, something they may not have considered when they would not have been asked this question.

Asking people for their level of trust is the equivalent of taking an intentional stance (Dennett, 1989). Put simply, one predicts behavior in such a case by ascribing to the human the possession of certain information and supposing it to be directed by certain goals. If someone does not trust the information presented, one may reasonably expect this person to display certain behavior, such as ignoring the information or discounting it. Although Dennett (1989) views the intentional stance as a stance taken by an external observer, Newell (1982/1990) takes a systems view and talks about the 'knowledge level' as an independent level above the symbol level (at which cognitive processes take place from 100 ms up to 10 s). Asking someone about their level of trust is a question directed at the

knowledge level rather than the symbol level. This is because the level of trust is an outcome of numerous cognitive processes. Credibility evaluation in the way I approach it takes longer than 10 s, hence should be viewed as a knowledge level process. The symbol level is appropriately addressed by employing measurement techniques such as reaction time or psychophysiological measurements. The knowledge level is appropriately addressed by employing measurement techniques such as think aloud or direct questionnaires. Therefore, I will ask for trust in the information using a single Likert-scale question (e.g., "How much trust do you have in this information?"). This question on the outcome will be followed by an open-ended question in which the participants can leave a motivation for their trust. This motivation can reveal several details of the credibility evaluation process, such as the features which are considered, or how such features impact trust. In Chapter 3, the think-aloud method gives an even greater insight into this process. By separately measuring the *process* and *outcome* of credibility evaluation, we can measure the influence of the independent variables (e.g., information quality, familiarity with the topic) on both levels. For instance, when manipulating the number of references, it will be possible to separately measure whether this has an influence on the way credibility is evaluated, and on the outcome variable: trust.

2.2 Participants

In most studies, students serve as a convenience sample for the experiments (n = 17), whereas only in a few of these cases, the authors specifically aim at studying students' behavior (Brand-Gruwel, Wopereis, & Vermetten, 2005; Julien & Barker, 2009; Lim, 2009; Head & Eisenberg, 2010). Students are of course not a representative sample of all Internet users; however, they represent a substantial group of Internet users (Lim, 2009; Head & Eisenberg, 2010). In other studies, a more general population of Internet users (often recruited online) was employed (n = 9). A few attempts have been made to create a representative sample (Dutton & Shepherd, 2006; Flanagin & Metzger, 2011). Finally, in a few studies (n = 4), renowned experts on the topic of the presented information were found.

We will also mostly employ students as participants in our experiments as a convenience sample. In Chapter 3, we will compare the evaluation behavior of users with varying levels of information skills, which is operationalized by selecting students from different stages of the educational process (similar to Brand-Gruwel, Vermetten, & Wopereis, 2005). This will have implications for the generalizability of the results, as students (especially at the

college level) are expected to have more than average information skills. It should thus be kept in mind that other, perhaps lesser educated Internet users are likely to have a less broad set of evaluation strategies. To cover a broader sample, in Chapter 2 a self-selected sample of Internet users will be tested, half with expertise in automotive engineering, and the other half without such expertise. In Chapter 6, a random sample of Internet users will be partaking in the experiment.

2.3 Task setting

The task that the participants had to perform also varies between the inspected studies in Table 1. In most studies (n = 16), the participants were directly asked to evaluate the credibility of the presented information (albeit termed slightly different in some studies; e.g., trustworthiness, reliability, information quality). In order to create a more natural task setting, some studies (n = 5) introduced an information problem that needed to be solved (i.e., searching for the answer to a given question). In three studies, the participants were asked to browse naturally as they would normally do; in five other studies, no task was provided at all, as these studies consisted merely of questionnaires.

The advantage of a strict task setting of credibility evaluation over a more natural task setting is that everything that a participant does can be related to credibility evaluation in a straightforward manner. However, this comes with the downside that the credibility of the information may be overemphasized in comparison to a real-life situation where credibility evaluation is merely a subtask of a larger task set, for instance writing a term paper for a course. This means that expecting exactly the same evaluation behavior in real-life may be somewhat optimistic. Rather, the behavior in the experiments can be seen as providing the full arsenal of evaluation strategies that a participant has. In real-life, it is likely that users select only a few of these strategies (Metzger, 2007; Walraven, Brand-Gruwel, & Boshuizen, 2009), although it cannot be ruled out that strategies other than those found in a lab environment are also applied.

Following this line of thought, the participants in the experiments discussed in this dissertation will be asked to evaluate credibility directly, rather than giving them a broader task set. Since Wikipedia will serve as the context in which trust is studied, this task is labeled the "Wikipedia Screening Task" (first introduced in Lucassen & Schraagen, 2010). The task instructions are straightforward: "Please rate the credibility of the presented article." The way to do this is not specified, which means that the participants are

encouraged to employ their own methods.

In some cases (Chapters 2 and 6), too much emphasis on the credibility of the information is undesirable. This will be addressed by either asking about the credibility only after the administration of the stimuli (Chapter 2), or boxing in the questions about credibility by other, unrelated questions (e.g., readability, information load).

2.4 Wikipedia

As mentioned above, in this dissertation credibility evaluation will be carried out on the information provided by the online open encyclopedia Wikipedia. The case of Wikipedia is an especially interesting one to study credibility evaluation. This is caused by something that I like to call the "Wikipedia-paradox". On the one hand, we know that overall, the information quality of Wikipedia is high (Giles, 2005; Chesney, 2006). However, because of the open-editing model behind Wikipedia (anyone can contribute, and modifications are directly visible online), users always need to be aware that the credibility of the information may not always be on the same level (Denning, Horning, Parnas, & Weinstein, 2005). Hence, Wikipedia users should always consider the credibility of the articles. Because of the quality variations between articles (and over time), the traditional approach to evaluate credibility by considering the source (Chaiken & Maheswaran, 1994; Sundar, 2008) is rendered invalid. Therefore, users should consider each single article they intend to use separately, based on other features than their source. In this dissertation we intend to find out what these features are, and how they are related to user characteristics.

Wikipedia has several features that make it a highly suitable environment for studying credibility evaluation in a controlled task setting. First, it has a standard quality rating system (Wikipedia: Version 1.0 editorial team, n.d.), allowing for selection of various articles in terms of quality. This facilitates repeatable selection of stimuli of comparable quality, and hence allows for stimulus control. Second, Wikipedia contains articles on so many topics that it is easy to match articles with user familiarity on a particular topic. By simply asking participants in advance what topics they are familiar with, an appropriate selection can be made. Third, the articles vary in length and may be selected in such a way that multiple articles may be evaluated in a single session, allowing for repeated measurements and more control over error variance. Fourth, it is well-known and the task of credibility evaluation has high face validity, specifically for a student user group.

Chapter 2

Factual Accuracy and Trust in Information: The Role of Expertise

In the past few decades, the task of judging the credibility of information has shifted from trained professionals (e.g., editors) to end users of information (e.g., casual Internet users). Lacking training in this task, it is highly relevant to research the behavior of these end users. In this article, we propose a new model of trust in information, in which trust judgments are dependent on three user characteristics: source experience, domain expertise, and information skills. Applying any of these three characteristics leads to different features of the information being used in trust judgments; namely source, semantic, and surface features (hence, the name 3S-model). An online experiment was performed to validate the 3S-model. In this experiment, Wikipedia articles of varying accuracy (semantic feature) were presented to Internet users. Trust judgments of domain experts on these articles were largely influenced by accuracy whereas trust judgments of novices remained mostly unchanged. Moreover, despite the influence of accuracy, the percentage of trusting participants, both experts and novices, was high in all conditions. Along with the rationales provided for such trust judgments, the outcome of the experiment largely supports the 3S-model, which can serve as a framework for future research on trust in information.

Published as Lucassen, T. & Schraagen, J. M. (2011). *Journal of the American Society for Information Science and Technology, 62,* p. 1232-1242.

1. Introduction

Since the 1980s, there has been a shift in responsibility for the verification of information credibility. Earlier, this task was mostly performed by professionals. Newspaper editors, for instance, used to decide which pieces of information were suitable for release to the general public. Credibility was one of the decisive factors for this decision, along with, for example, relevance to the public and readability. Nowadays, the task of distinguishing credible information from less credible information often lies with the end user of the information (Flanagin & Metzger, 2007). The introduction of the World Wide Web (and especially Web 2.0) has resulted in a much larger range of information suppliers than before, for which expert evaluations of credibility are often not available. Online information is not less credible, per se, but users should be aware of the possibility of encountering low-quality information. A good example is Wikipedia: Research has shown that its information quality is overall very high (e.g., Giles, 2005; Rajagopalan et al., 2010), but the open-editing model combined with the tremendous number of articles (>3.4 million; Statistics, n.d.) requires users to always be aware of the risk of low-quality information (Denning, Horning, Parnas, & Weinstein, 2005).

A highly relevant topic for research is how lay people cope with the varying credibility of information. While they need to make assessments of credibility, they typically are not trained for this task as are professionals. It is suggested in the existing literature that individual differences among users influence trust assessment behavior. In this study, we attempt to explain these differences in terms of user characteristics, particularly focusing on trust in information of Internet users with varying levels of expertise on the topic at hand. This relationship between domain expertise of the user and trust judgments is especially new in the field of information credibility and trust research.

In this article, we first discuss the concept of trust, of which no consensus has been reached by researchers in the various relevant fields. Second, we propose a new model of trust in information. We use this model to predict that various characteristics of a user lead him or her to employ different features of the information to judge its credibility. We then continue to discuss in detail three types of relevant user characteristics: domain expertise, information skills, and source experience. Our hypotheses aim at validating the proposed model. After this, our method using online questionnaires featuring Wikipedia articles with manipulated accuracy is introduced to test the hypotheses. Finally, the results are presented and discussed.

1.1 Trust

The concept of trust has been studied in various ways in the literature. Kelton, Fleischmann, and Wallace (2008) distinguished four levels of trust: *individual* (an aspect of personality), *interpersonal* (one actor trusting another), *relational* (an emergent property of a mutual relationship), and *societal* (a characteristic of a whole society). The most common approach of studying trust is at the interpersonal level, concerning a one-way tie between a trustor (someone who trusts) and a trustee (someone who is trusted). An often-used definition of trust at this level was given by Mayer, Davis, and Schoorman (1995):

> *The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (p. 712)*

Trustors assess trustees to determine the degree to which the trustees can be trusted. These assessments often include estimating various characteristics of the trustee, deemed relevant to trustworthiness by the trustor (trust antecedents). These typically include factors such as perceived competence, intentions, and openness.

According to Kelton et al. (2008), interpersonal trust also is the appropriate level to apply to the study of trust in information because information is produced by an author (trustee) and communicated over a certain channel to a receiver (trustor). Assessing trust in the information thus can be seen as assessing trust in the author. However, next to the assessment of characteristics of the author, assessing trust in information also may include characteristics (features) of the information itself. This approach seems especially useful when the author of the information is unknown or when a piece of information has multiple authors. An example of such a case is Wikipedia, where multiple, often anonymous authors contribute to one article. In such situations, the assessment of characteristics of the author(s) may become overly complex or even impossible. Alexander and Tate (1999) identified five criteria that always should be considered when assessing trust in information: accuracy, authority, objectivity, currency, and coverage. Cues of at least four of these criteria (all but authority) also may be found in the information itself, without knowing the identity of the author.

A term often used interchangeably with (information) trust is credibility; however, there is a slight difference. Fogg and Tseng (1999) summarized this difference as credibility meaning believability and as trust meaning dependability. Credibility can be described

as perceived information quality, or the assessment of information quality by a user. Credibility is mostly seen as consisting of two key elements: trustworthiness (well-intentioned) and expertise (knowledgeable). Trust, however, also introduces the notion of willingness to depend on the credibility of information. This dependency involves a certain risk that someone takes by using the information (Kelton et al., 2008).

In the remainder of this article, we refer to "trust" as a property of the information user. Credibility is used as the aspect of information that is being considered when judging trust.

A model of online trust proposed by Corritore, Kracher, and Wiedenbeck (2003) has shed more light on the relationship between trust and credibility. Two factors influencing trust were identified in their model: external factors and individual perception. External factors can influence the perception of trust, which in turn is composed of three factors: credibility, ease of use, and risk.

Kelton et al. (2008) proposed an integrated model of trust in information. According to this model, trust also may stem from other factors than the assessment of trustworthiness, such as the disposition to the information, relevance of the information, and recommendations. Personal factors, such as confidence and willingness to trust, also may contribute. This suggests that users with varying (personal) characteristics may judge the same information very differently.

The unifying framework of credibility assessment, as proposed by Hilligoss and Rieh (2008), also acknowledges the influence of personal characteristics on judgment behavior. Three levels of credibility assessment between the information seeker and information object were distinguished in interviews with undergraduate students. First, the construct level describes the users' personal definition of credibility. This may include concepts such as truthfulness, believability, and trustworthiness. The definition of the user may deviate from the definition given by Fogg and Tseng (1999) since mental models of the construct may vary exceptionally between users due to, for instance, differences in age, education, or intelligence. The second level is labeled *heuristics* by the authors and refers to general rules-of-thumb used to estimate credibility. These heuristics include media-related and source-related heuristics. The third level concerns actual interaction with the information, which can be split into content and peripheral cues from the information itself as well as from its source.

The content and peripheral cues in the interaction level of the framework proposed by Hilligoss and Rieh (2008) is similar to the distinction between heuristic and systematic evaluation. Metzger (2007) also made this distinction in her dual-processing model of website credibility assessment. This model is strongly based on the dual-processing theory of Chaiken (1980) and predicts the type of assessment done by a user, depending on the motivation and ability to evaluate. Metzger defined heuristic evaluation as using superficial cues and systematic evaluation as constituting a thorough evaluation of a website's credibility.

Motivation comes from the "consequentiality of receiving low-quality, unreliable, or inaccurate information online" (Metzger, 2007, p. 2087). Motivation thus can vary, as consequences of low-quality information might differ between tasks. For tasks with low importance (e.g., personal entertainment purposes), consequences of poor information could be very limited whereas tasks of higher importance (e.g., searching information for a school assignment) can have more serious consequences (e.g., a low grade). Motivation thus can be interpreted as the importance of credible information. When the user is not motivated, no evaluation is done at all or a heuristic evaluation is done. When the user is motivated to evaluate, however, the type of evaluation depends on the ability of the user. Ability is linked to "the users' knowledge about how to evaluate online information" (Metzger, 2007, p. 2087). These skills can be taught to users in information skills education. If a user has the ability to evaluate, a systematic/central evaluation is done; otherwise, a heuristic/peripheral evaluation is done.

A different approach was taken by Fogg (2003). His prominence-interpretation theory predicts the impact of various noticeable elements in a piece of information on a credibility assessment. Prominence refers to the likelihood that an element is being noticed by the user. This is multiplied by interpretation, which indicates the value or meaning people assign to this element. The result is the credibility impact of the element under evaluation.

Metzger's (2007) model mainly considers aspects of users' motivation and ability whereas Fogg's (2003) theory concerns the information itself without identifying aspects of the user, which may lead to different prominence or interpretation of elements. Combining the predictions of both models, one can expect that the influence of various elements in a piece of information is based on specific characteristics of a user. Metzger predicted that the type of evaluation is dependent on the ability of the user, but various levels of ability also could lead to other elements being prominent in a piece of information. An example

is the element of "references" in an article. For academic students, this is a very prominent element (Lucassen& Schraagen, 2010); however, younger school children are probably not (yet) familiar with the concept of referencing or its importance (Walraven, Brand-Gruwel, & Boshuizen, 2009).

This aspect of a user's ability that Metzger (2007) described in her dual-process model is quite general. We propose to distinguish two types of expertise on the topic at hand: (generic) information skills and domain expertise. Both have the potential to influence a user's ability to assess credibility. When a piece of information is within the area of users' expertise, different elements are likely to be prominent as compared to information outside their area of expertise. Using elements such as accuracy, completeness, or neutrality requires knowledge of the topic at hand, which only users with a certain level of domain expertise have. However, other elements, such as the length of a piece of information or the number of references do not necessarily require domain expertise.

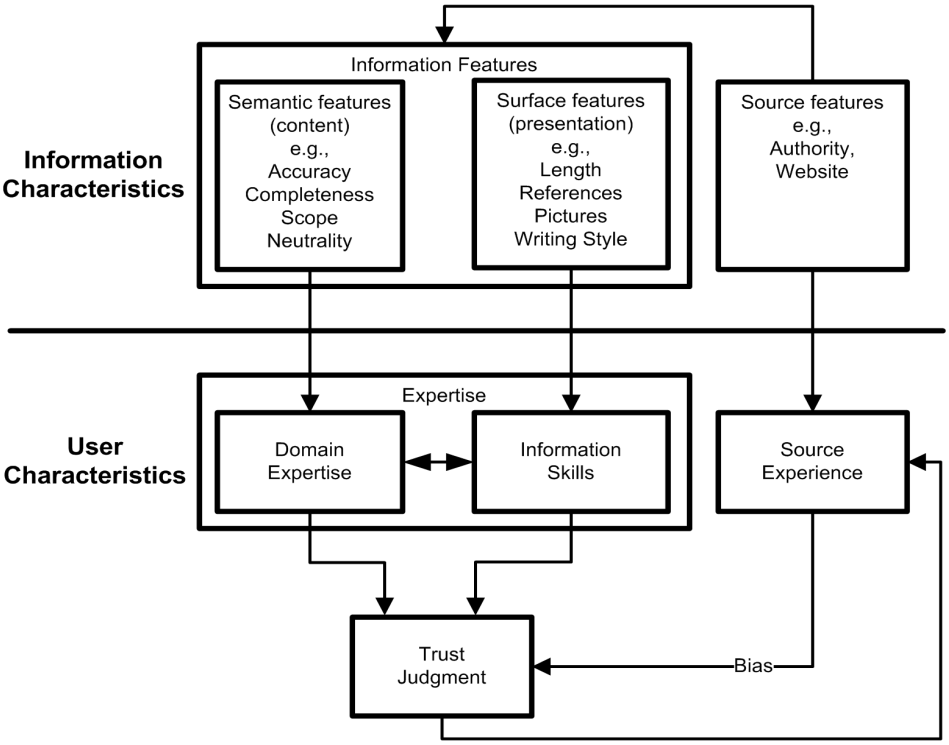

Figure 1: The proposed 3S-model of information trust.

In this article, a new model of trust in information is proposed, as shown in Figure 1. In this model, we predict that trust judgments of a user specifically depend on the two aforementioned user characteristics: information skills and domain expertise. Based on prominence-interpretation theory (Fogg, 2003), these characteristics lead to different features in the information being used in trust judgments. Furthermore, users may alternatively choose to rely on their earlier experiences with a particular source instead of actively assessing various features of a piece of information. In this model, we have tried to add more detail to the trust behavior of users than do current models by considering characteristics of both the user and information.

We name the proposed model the *3S-model*. The three Ss stand for *semantics*, *surface*, and *source* features of information, as well as for the three different strategies users may take when judging credibility of information. We discuss these three main elements of the proposed model in detail in the following sections.

1.2 Domain Expertise

Expertise has a long history in psychological research. It is well-known that experts approach problems within their domain of expertise differently than do novices. Whereas novices are known to think about problems in a concrete manner, focusing on surface characteristics, experts tend to form abstract representations, focusing on the underlying principles of a problem. For example, Chi, Feltovich, and Glaser (1981) found evidence for this difference by presenting physics problems to both experts and novices. The participants in this experiment were asked to categorize these problems into groups based on similarity of solution. Virtually no overlap was seen between the categories introduced by novices and experts. Novices tended to sort the problems according to surface features, such as the presence of a block on an inclined plane in the description of the problem. In contrast, experts generally categorized the problems into groups based on the underlying physics principles that could be applied to solve the problem.

Adelson (1984) used the same distinction between experts and novices to create a situation in which novices could actually outperform experts. Undergraduate students and teaching fellows were considered novices and experts, respectively, in the domain of computer programming. Two conditions were introduced; in the first condition, a concrete representation of a computer program was given (concerning how the program works), after which a concrete question was asked. In the second condition, both the

representation and the question were abstract (concerning what the program does). The first condition should better suit novices whereas the second condition should suit experts. This hypothesis was confirmed by the measured task performance; experts were better in answering abstract questions, whereas novices answered more concrete questions correctly.

When domain experts and novices are asked to judge information credibility, similar differences to those found by Chi et al. (1981) and Adelson (1984) can be expected. When experts judge information within their area of expertise, they are able to assess the *content* on several aspects such as accuracy, neutrality, or completeness. Novices are less able to do this due to their lack of knowledge about the topic; they mainly have to rely on the assessment of *surface* characteristics.

Domain familiarity can be seen as a weaker form of domain expertise. In a think-aloud study by Lucassen and Schraagen (2010), familiarity with the topic was varied for participants judging credibility. While no significant difference was found in the distribution of information features used, post-hoc inspection of the data showed that correctness of the information was mentioned almost solely by participants familiar with the topic. Correctness (or accuracy) of the information thus may be an important factor for trust in information, which can predominantly be judged when the user has a sufficient level of domain expertise.

1.3 Information Skills

As noted earlier, users may judge other aspects than the semantics of a text as well, such as surface features. Assessing such features does not necessarily require domain expertise; other skills are needed to identify which features are relevant to credibility. These skills can be seen as a subset of information skills. A common definition of this is "the ability to recognize when information is needed and the ability to locate, evaluate, and use effectively the needed information" (American Library Association Presidential Committee on Information Literacy, 1989). In this study, we focus on the evaluation aspect as this includes evaluation of credibility or trust. We interpret information skills as generic skills, which require no expertise in the domain of the information.

Users with varying levels of information skills approach information in different ways. Brand-Gruwel, Wopereis, and Vermetten (2005) investigated information problem

solving by information experts (doctoral students) and information novices (psychology freshmen). The task of information problem solving was decomposed in problem definition, searching, scanning, processing, and organization of the information, all guided by a regulation process. Judging information (including credibility) is done in the information scanning and processing stages. The first stage can be seen as heuristically scanning the information whereas the latter stage involves in-depth systematic processing of the information. They found that experts put significantly more effort in the processing stage than do novices. Experts also seem to judge scanned information more often, although a difference was found only at the 10% significance level. These findings indicate differences in behavior between experts and novices in judging information, especially since their behavior was largely similar in most other stages of information problem solving.

Brand-Gruwel et al. (2005) further showed a difference in the amount of effort information experts and novices put into the processing of information. However, qualitative differences also can be expected. Walraven et al. (2009) for instance, showed that in group discussions by people with limited training in information skills (high-school students), many factors relevant to trust or credibility are not mentioned. Examples are objectivity and whether information comes from a primary or secondary source, but the notion of references also was mentioned only once in eight group discussions.

Lucassen and Schraagen (2010) showed that for college students, several textual features, references, and the presence of pictures were important noncontent features when judging credibility of Wikipedia articles. The differences between the importance of references for high-school students and college students can be attributed to differences in information skills. Hence, people with varying information skills can be expected to differently assess credibility of information.

We do not suggest that the strategies of employing domain expertise or information skills to form a trust judgment are mutually exclusive. Instead, we expect that for various users, strategies vary in their impact on the trust judgment. For instance, domain experts are likely to base their judgment primarily on factual accuracy whereas people with advanced information skills (e.g., information specialists, doctoral students) are likely to mostly bring to bear their information skills in their judgments when the topic at hand is out of their domain. However, it is not expected that domain experts will no longer notice surface features or that information specialists no longer notice the semantics; their

domain expertise or information skills may only render certain features more prominent than others.

Furthermore, we expect that both types of user expertise interact. Consider, for example, the quality of references. Domain experts will know which journals are considered the best in their field. This knowledge can aid the information skills of the user and improve the trust judgment.

1.4 Source Experience

An alternative strategy to form a trust judgment also is introduced in the 3S-model. Instead of *actively* assessing content or surface features, the user may *passively* rely on earlier experiences with the source of the information. This behavior also was identified by Hilligoss and Rieh (2008) in the heuristics level of credibility assessment (source-related heuristics). Following this strategy, it is possible that the influence of domain expertise or information skills (and thus the corresponding features in the information) is diminished or even ruled out when a user has a lot of positive (or negative) experiences with a particular source. In this case, a user will no longer feel the need to actively judge the credibility of the information, which is similar to the prediction of Metzger (2007) that the lack of motivation leads to no assessment or a heuristic assessment.

When a trust judgment is formed following any of the three proposed strategies, this new experience is added to the preexisting experience with the source. This feedback connection also is present in the integrated model of trust in information by Kelton et al. (2008).

1.5 Heuristic versus Systematic Processing

Using one's experience with the source of information to judge credibility can be considered highly heuristic behavior. However, semantic and surface features can be evaluated heuristically or systematically. While some of the features listed as examples of surface features at first might seem to facilitate heuristic processing (e.g., the length of a text), surface features also can be processed systematically. An example is assessing the quality of the references: Doing this requires an effortful evaluation of each reference. The same is true for the assessment of content features: At first, this may seem to require systematic processing, but the process of comparing presented information with own knowledge can be considered recognition, which according to the RPD model (Klein, Calderwood,&

Clinton-Cirocco, 1986) does not require systematic comparison of information. On the other hand, when a presented statement is just outside of the area of expertise, its validity might still be checked by bringing to bear the knowledge an expert possesses, which is typically a systematic process (resulting in the phenomenon of "fractionated expertise," described by Kahneman & Klein, 2009, p. 522).

However, we argue that assessing trust in information always will contain a certain degree of heuristics. Consider someone who systematically evaluates every single element relevant for trust in a piece of information. By doing this, the risk of using poor information is eliminated, which in itself is an important aspect of trust (Fogg & Tseng, 1999; Kelton et al., 2008). This means that trust is no longer necessary because the user has complete certainty of the credibility of the information. However, complete certainty is impossible; hence, trust assessments are always heuristic to a certain degree. Grabner-Krauter and Kaluscha (2003) also identified this in their proposition that trust and information search (systematic processing) are alternative mechanisms to absorb uncertainty. This is needed because situations are generally too complex to incorporate all relevant factors.

1.6 Hypotheses

In this study, we attempt to find empirical evidence for the validity of our proposed model, mainly focusing on the concept of domain expertise. We asked Internet users with varying expertise in one particular area (automotive engineering) to assess the credibility of Wikipedia articles on this topic. The factual accuracy of the articles was manipulated, ranging from original quality to articles containing factual errors in half of the treated concepts as well as in the topic definition. According to the proposed model, lower accuracy should affect trust judgments of users with domain expertise. This leads to the first hypothesis:

H1: Decreases in factual accuracy have a negative impact on trust in information of domain experts.

We hypothesize that users with little domain expertise are less able to focus on content features to assess credibility. This would mean that manipulating accuracy does not influence the trust judgments of these users, which leads to the second hypothesis:

H2: Decreases in factual accuracy have no impact on trust in information of novices.

These hypotheses are based on the expectation that domain experts and novices will use different cues from the article in their assessments. A substantial number of these cues can be made explicit by asking users for their rationales for their judgments (We acknowledge that some of this knowledge may be tacit and not open to verbalization.) According to the 3S-model, this leads to the final two hypotheses:

H3: Novices use surface and source features more than semantic features in trust judgments.

H4: Experts use semantic features to a larger extent than do novices in their trust judgments.

Note that the expectation that experts will use their domain expertise does not give reason to assume that they will no longer use surface features to assess credibility. This could be the case when domain experts with very limited information skills are assessing credibility, but testing such hypotheses is beyond the scope of this study.

2. Method

2.1 Participants

Since nearly every car brand (and model) has its own online forum with numerous members, automotive engineering was used as the domain of expertise for this experiment to easily recruit a large number of participants. Experts were mainly active at car enthusiasts' forums whereas novices were recruited mainly from other, general-purpose forums. Invitations for participation were posted on these forums, containing a link which led them to an online questionnaire. A total of 657 participants took part in the experiment (70.0% male). The average age was 27.7 years (SD = 10.0). We identified 317 experts and 340 novices (Definitions used for "expert" and "novice" are discussed later.) Since all participants were Dutch or Belgian (Flemish), the experiment was performed in Dutch, using articles from the Dutch Wikipedia.

2.2 Task and Procedure

The experiment was implemented in the form of an online questionnaire. When it was opened, an explanation of the experiment was provided, including an indication of its duration ("a few minutes") and the number of questions (n=8). Participants were told that

they would be asked for their opinion on one Wikipedia article, without specifying what aspects of the article their opinion should be about. By doing this, we made sure that the participants were not primed to specifically focus on the credibility of the article but to approach the article in a more natural manner. After reading the instructions, participants were asked to provide some general demographic information such as gender, age, and education level. On this page, they also were asked whether they worked in the automotive industry and whether they considered cars to be a hobby.

On the subsequent page, a Wikipedia article was presented. Three different articles were used in the experiment to account for potential influences of characteristics specific for one particular article (e.g., a very lengthy article or an unusually high number of images). The topics used were "V-type engine", "Boxer-type engine", and "Twin turbo". The articles were selected to be of similar appearance (e.g., length, presence of images) and topic (car engines). Each participant viewed only one randomly selected article. It was not possible to click on the links in the article since a full-page screenshot (using WebShot; Moinvaziri, 2012) of the actual article was presented.

After the participants indicated that they had finished reading the article, they were asked whether they trusted it by means of a yes/no question. Next to this, a rationale for their judgment could be provided optionally. The trust question and the rationale were presented on a separate page from the Wikipedia article. To prevent multiple page views when answering the questions, it was not possible to go back to the article once the participants indicated that they had finished reading the article. The participants were made aware of this in the instructions.

To ensure that participants could fill in the questionnaire only once, IP addresses were registered, and a cookie was saved on the participants' computers. Due to the technical limitations of online questionnaires, it could not be ensured that the participants cross-checked information with other websites or visited the original page on the Dutch Wikipedia; however, none of the rationales indicated such behavior. Furthermore, we do not expect that such behavior would interfere with the goals of this study.

2.3 Independent Variables

2.3.1 Expertise

This variable was assessed using two questions. Participants who indicated that they

worked in the automotive industry or who considered cars as a hobby were considered experts; otherwise, they were considered novices. The participants were not asked directly whether they were experts in the domain because we expected that this might lead them to read the article in a different way (e.g., especially focusing on their domain expertise). We acknowledge that this strategy of distinguishing experts from novices does not guarantee that our expert participants were absolute domain experts. However, we expect the differences in domain familiarity and expertise between our expert and novice participants to be sufficient for the purpose of this study.

2.3.2 Factual accuracy

This variable was manipulated by adding factual errors to the article. First, the number of concepts treated in each article was counted. Then, the facts in a predefined percentage of concepts were altered in such a manner that no inconsistencies within the article were created. Possibly due to the descriptive encyclopedic character of Wikipedia articles, there were only a few links between the concepts in one article. This means that single facts could be altered while maintaining internal consistency.

Furthermore, the facts were altered to be the opposite of the actual fact, or at least very different from it. By doing so, the presented facts were clearly incorrect. An example of an altered fact in the article on the "V-shaped engine" is the following sentence: "V-shaped engines are mostly applied in vehicles in which space is not an issue. By placing the cylinders diagonally, a V-engine takes more space than an inline or boxer engine with the same capacity."[1] Originally, the article correctly stated that these engines are applied when space is an issue because they take up less space.

The articles used were not very extensive (~600 words) and provided a brief introduction on the topic rather than an in-depth discussion. Therefore, we could assume that people with a reasonable level of domain expertise would be able to detect at least some of the errors introduced.

The manipulation was validated by showing all original and manipulated statements of each article side by side to two independent domain experts (garage owners). A substantial degree of intersubjective truth about the correctness of the statements was reached since they were able to identify the correct and incorrect statements with an accuracy

1 Note that this is a translation of the original sentence; the articles used in the experiment were in Dutch.

of 92.3%. Only two statements were not correctly identified by the domain experts. The first statement was on the English term for a boxer engine with four cylinders (flat-four), which was incorrectly identified by both garage owners. This error likely can be attributed to a lack of proficiency in the English language. The second statement was on the angle between the cylinders and the crankshaft in a V-shaped engine, which was incorrectly identified by one of the garage owners. It is most likely that this statement was misread since this domain expert can be assumed to be highly familiar with V-engines.

The following conditions were used in the experiment:

- Original article, not manipulated
- Errors in 25% of the concepts
- Errors in 50% of the concepts
- Errors in 50% of the concepts and an error in the topic definition.

The definition of the topic is given in the first sentence of each article. It is presumably more important than the other concepts because it introduces the main concept of the article and helps to get a grasp of the subject of the article. For example, the correct definition of the "Biturbo" article is "A biturbo or twin-turbo is an internal combustion engine fitted with two turbos." The manipulated definition stated that biturbo engines are diesel engines. The conditions were randomly assigned to the participants.

2.4 Dependent Variables

2.4.1 Trust judgment

The percentage of the participants trusting the information in the article in each condition was measured by the percentage of positive answers to the question "Do you trust the information in the article?" A dichotomous scale was used because each participant assessed only one article. More detailed scales (e.g., a 7-point Likert scale) were considered less useful because participants could not compare articles.

2.4.2 Rationale for the trust judgment

The (optional) rationales for the judgments of participants were categorized into the three strategies proposed in the 3S-model: rationales based on surface features, semantic features, and source features. Rationales containing comments on multiple features were categorized according to the dominant feature in the rationale. Rationales that could not

be categorized into one of these types were classified as "other". Two experimenters both analyzed 60% of the data; Cohen's κ was calculated for the overlapping 20%. The resulting value of 0.799 indicates a substantial agreement. A qualitative analysis of the disagreements between annotators revealed that most of them were the result of interpretation differences of the most dominant feature in rationales with multiple features.

3. Results

3.1 Trust Judgments

Table 1 and Figure 2 show the percentages of experts and novices trusting the information in the articles in all conditions.

The percentage of experts trusting the information decreased when the factual accuracy of the articles was manipulated, $\chi^2(3) = 7.81$, $p = .05$, supporting H1. For novices, no difference was found, $\chi^2(3) = 3.69$, $p = .30$. This supports H2.

Visual inspection of Figure 2 shows an unexpected "dip" for trust of novices in the 25% condition. Post hoc analysis showed that the number of novices trusting the information in this condition was almost significantly lower, $\chi^2(1) = 3.37$, $p = .066$, than that in the other conditions. Analysis of the novices' trust judgments in this condition showed no significant differences between the three articles used, $\chi^2(2) = 1.71$, $p = .43$. Subsequently, a quantitative and qualitative inspection of the rationales of novices in this condition also revealed no differences compared to those in the other conditions, $\chi^2(1) = 0.15$, $p = .70$. Furthermore, content of the articles in this condition was examined post hoc. No unexpected irregularities in terms of, for instance, internal consistency were found.

3.2 Rationales for the Trust Judgments

A total of 520 participants (79%) gave a short rationale for their trust judgments. Table 2 gives an overview of how these rationales were divided into the categories of source

Table 1: Percentages of participants trusting information in the article for varying manipulation levels. The exact numbers of experts and novices trusting the information in each condition is given in parentheses.

	Original article	25 % errors	50% errors	50% errors + definition
Experts	82.2% (60 of 73)	79.3% (65 of 82)	71.3% (57 of 79)	64.6% (53 of 82)
Novices	69.8% (60 of 86)	59.5% (47 of 79)	69.0% (60 of 87)	72.7% (64 of 88)

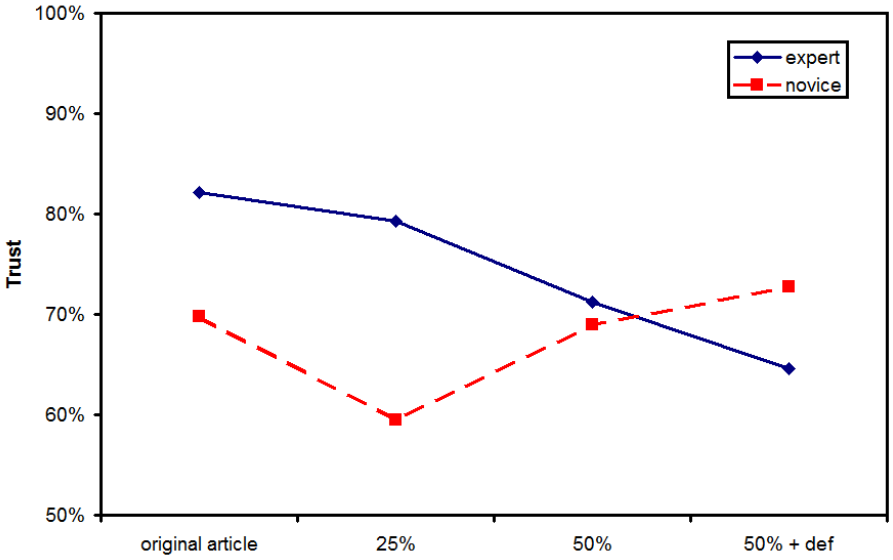

Figure 2: Percentages of experts and novices trusting the information in each condition.

Table 2: Feature categories used in trustworthiness assessments by experts and novices. Categories with a significantly higher number of rationales by experts or novices are marked by an asterisk.

	Experts	Novices
Source features (source experience)	25.2% (*n*=60)	33.7% (*n*=95)*
Semantic features (domain expertise)	38.2% (*n*=91)*	6.7% (*n*=19)
Surface features (information skills)	32.8% (*n*=78)	47.9% (*n*=135)*
Other motivations	3.8% (*n*=9)	11.7% (*n*=33)

features, semantic features, surface features, and other rationales.

Rationales in which the source of the information (Wikipedia) was mentioned were classified as source experience. Examples of these are "I don't trust the information, because it's from Wikipedia, and anyone could have put it on there, without having any knowledge on the topic" and "This is from Wikipedia, which is mostly quite accurate." Rationales regarding factual accuracy or the preexisting knowledge of the participant were categorized as domain expertise. For instance, rationales such as "This fits with my own knowledge" and "I found some factual errors in the article" were classified as reflecting domain expertise, including rationales in which specific errors in the article were mentioned (e.g., "There never was a Boxer engine in the Alfa Romeo 156"; "This is wrong, larger turbos do not work better at lower speeds"). When surface features of the article were mentioned, the rationales were categorized as information skills. Examples

of these are "This looks well-written" and "Poor language, few references supporting the propositions." All rationales that could not be categorized in one of these three categories were marked as "other" rationales and excluded from further analysis due to their diverse character and low percentage (8%).

The distribution of features used by novices was different from a distribution expected by chance, $\chi^2(2) = 83.66$, $p < .001$. Considering the low percentage of semantic features (6.7%), this supports H3, which predicted that novices would largely focus on source and surface features. The three proposed categories covered 88.3% of all rationales of novices.

The distribution of cues across the three categories (see Table 2) was different for experts and novices, $\chi^2(2) = 69.57$, $p < .001$. Experts used semantic features from the information more than did novices to underpin their trust judgments, $\chi^2(1) = 69.41$, $p < .001$, supporting H4. Moreover, novices relied more on the use of surface features, $\chi^2(1) = 19.62$, $p < .001$, and source features, $\chi^2(1) = 7.78$, $p < .01$, than did experts. The three strategies proposed by the 3S-model covered 96.2% of all rationales of experts.

Post hoc analysis showed that the source of the information was used as a rationale to both trust and not trust the information. No significant difference was found between positive and negative use of this rationale, $\chi^2(1) = 1.09$, $p = .30$. Furthermore, experts and novices did not differ in their ratios of positive to negative "source" rationales, $\chi^2(1) = 0.68$, $p = .41$.

4. Discussion

The trust judgments of participants in the various conditions of this study largely support the proposed 3S-model. We found that trust of domain experts was influenced by the accuracy of the presented information. This was expected because accuracy is a key aspect of the semantic (content) features of information.

According to the proposed model, evaluating these features requires a degree of domain expertise. In contrast to domain experts, novices' trust remained approximately the same for the various accuracy levels. The 3S-model can be used to predict that their lack of domain expertise leads them to mainly assess source and surface features, which were kept constant in this study. Hence, their trust judgments were not influenced by the manipulation of factual accuracy.

A second observation regarding trust of experts and novices is that the latter group had

less trust in the information in articles of original or slightly manipulated quality. This result replicates the finding by Chesney (2006) that domain experts value Wikipedia articles as more credible than do novices. His conclusion was that Wikipedia articles are very accurate, supporting Wikipedia as a reliable information source. We hypothesize that experts have the advantage of recognizing the presented facts in the article. This gives them a very strong sense of confidence in the information since their preexisting knowledge contains the same facts. Since novices mostly lack preexisting knowledge, they do not get this sense of confirmation. When the accuracy was severely manipulated, trust of experts was similar to trust of novices, or even less. Under these conditions, the recognition of facts by experts is replaced by the recognition of errors, significantly decreasing trust.

A second explanation for the low trust of novices in comparison to experts is that "distrusting" novices are aware of their own limited abilities to judge the credibility. To avoid potential problems as a consequence of the use of poor information, novices may be highly skeptical of the information they encounter. It is trusted only when they are highly confident of the credibility. This behavior might protect them from potentially poor information, but it also may keep them from using high-quality, credible information on unfamiliar topics. This hypothesis is supported by some statements expressed by novices, such as "I don't understand the information. A lot of terms are used which I don't understand and which thus can be wrong"; and "I know nothing about this topic, so I am not 100% confident that this information is true."

An important observation in the trust judgments of experts is that despite the negative influence of diminished factual accuracy on trust, the majority of the experts still trust the information. This was observed even in the condition with the highest percentage of errors (64.6% trust of experts). This observation can be explained in several ways.

First, the experts did not exclusively use their domain expertise in the assessments. In fact, numerous rationales still referred to aspects of the source or surface features of the information. The usage of these features does not lead to variations in trust between the accuracy conditions, as source features and surface features were kept constant.

Second, the participants were intentionally not made aware beforehand that they would be asked to judge the credibility of the article. Instead, they were instructed that they would be asked their opinion, without specifying what their opinion should be about. As this was done to stimulate natural behavior on Wikipedia, it also might mean that concepts such as trust or factual accuracy were not salient to the participants. They might instead have

been paying attention to other aspects of the information, such as the information load or entertainment value.

Moreover, we may expect a low motivation from the participants in our experiment since they had no personal benefit in performing well. According to Metzger (2007), this leads them to perform no evaluation at all or a heuristic evaluation. Answering the questions after viewing the article required them to do an evaluation, which thus was likely to be of a heuristic nature. Consider the large percentage of evaluations by experts based on semantic features. These evaluations might not have gone beyond swiftly skimming though the article, recognizing some of the presented facts, and inferring that all information in the article is credible based on the recognized facts. This hypothesis is supported by the fact that even in the condition with the worst factual accuracy, 50% of the treated concepts did not contain factual errors. By reading swiftly, the experts might have missed them.

A final explanation of the high number of experts trusting the information in conditions with heavily manipulated accuracy concerns their level of expertise. Our expert participants indicated that they either worked in the car industry or that they were car enthusiasts, but this does not necessarily mean that they were real experts on car engines. While the presented information was aimed at the general public rather than domain experts, this does not exclude the possibility that some of the facts were actually unknown to some of our expert participants. To find out whether our participants were able to find errors in the articles, a post hoc analysis of the rationales of experts was performed. This showed that 62% of the introduced errors were explicitly mentioned by at least one participant. However, more errors may have been found, as the participants often stated that they only found factual errors, without specifying them.

The rationales for the trust judgments of the participants in our study provide additional proof for the validity of the model. We observed that experts mainly try to bring to bear their domain knowledge with their judgments. However, this did not rule out the utilization of their source experience or information skills. In fact, numerous rationales still referred to aspects of the source or surface characteristics of the information. Following this observation, note that the use of domain expertise, information skills, and source experience are not mutually exclusive in trust judgments; instead, a combination of these features is employed. The impact of each feature in one of the three categories depends on the characteristics of each particular user and piece of information. In this study, we observed that domain experts in automotive engineering largely used their domain

expertise. Accordingly, we also predict that, for example, information specialists (e.g., librarians) will largely use their information skills and will therefore be largely influenced by various surface features (e.g., references).

Novices rarely mentioned semantic features in their rationales. This was expected because their domain expertise is at most very limited, if not completely absent. Novices mainly seem to compensate for their lack of domain expertise by assessing surface features of the information. This was reflected by a higher percentage of surface features in rationales of novices than in that of experts. Moreover, the use of source features by novices also exceeded experts' use of these features. Source experience and information skills do not require domain expertise on the topic and are thus highly accessible to novices.

As predicted by the source experience component of the 3S-model, the presented information was frequently dismissed simply because it came from Wikipedia. Experience with this particular source was clearly negative in these cases. Some participants mentioned this explicitly, whereas others referred to the underlying principles of Wikipedia (e.g., open-editing model, multiple authors). Remarkably, the source of the information also was used as a reason to trust the information, possibly because of earlier positive experiences with the website. In the case of Wikipedia, this is likely because the overall quality of Wikipedia is quite high (Giles, 2005). People who expressed this rationale did not assess content or surface features but directly gave their trust judgment based on earlier experiences with Wikipedia. These observations of the source of the information leading to a trust judgment in which the actual content of the information was not considered confirm the biasing influence of this strategy.

The limited domain expertise (of novices), which is expected in information search behavior (Lim, 2009), and limited information skills (of both novices and experts; Walraven et al., 2009) might have been the cause of the observation that users solely rely on previous experiences with the source. In most cases, this is not a problem because of the high overall information quality, but in cases when the quality of an article is disputed, users are unlikely to detect this following this strategy. Examples of such cases are vandalism or disputed neutrality (Denning et al., 2005). The high number of rationales in which the source of the information was mentioned also is a good indicator of how Wikipedia is trusted blindly by many and carefully avoided by others.

4.1 Limitations

The experiment performed in this study has brought some confirmation of the validity of the 3S-model; however, a few limitations should be kept in mind. The participants in the experiment were recruited from online forums. While this is a great strategy to obtain a high number of participants, accountability is low. For instance, experts were distinguished from novices only on the basis of two questions prior to the experiment. We have no reason to assume misbehavior, but novices also could have posed as experts by answering these questions in a particular way. Moreover, as stated earlier, experts might have been car enthusiasts without being domain experts in car engines. This leads to the limitation that we cannot be absolutely positive that our expert participants can be considered actual domain experts. However, their level of expertise proved to be adequate for the purpose of the experiment.

Each domain of expertise will have its own specifics concerning evaluation behavior. In this experiment, we have shown that for this setting in automotive engineering, the 3S-model seems valid. However, other domains may have different specifics, potentially leading to different behavior. Examples are differences in the consequences of poor information, controversy within the domain, or education level of domain experts. The 3S-model should be investigated using other areas of expertise.

In this research, the Dutch Wikipedia has been used as a case study to provide a familiar source of information, used by numerous people. However, lots of characteristics of the information, such as the layout or the open-editing model behind it, are very specific for Wikipedia. The 3S-model should be tested on different information sources in different contexts. Both online and offline sources should be considered. User scenarios other than handling encyclopedic information also could be applied. When, for instance, health information is considered, motivation could be much higher because of the potentially high impact of the negative consequences of poor information. A second domain in which credible information is vital is the military.

The rationales for trust judgments of the participants provided valuable insights into their behavior; however, note that these could be provided optionally, and not all participants did so. This means that these results may not apply to the entire sample in this experiment.

4.2 Future Research

More empirical research into the 3S-model is necessary. Fine-grained insights into the behavior of users following the three proposed strategies and the elements of information which correspond to these strategies should be attained. This could be achieved, for instance, by conducting think-aloud experiments. Furthermore, the performed experiment focused on the manipulation of features in one of the three strategies (domain expertise). Although we have shown the employment of all three proposed strategies, future experiments also should focus on manipulating features in the other strategies (source experience and information skills).

The lack of a difference between novices' trust in credible and less credible information is an important area to research. While this study has demonstrated this problem, more detailed (within-subject) studies should further investigate it, as this leads to novices not trusting credible information as well as novices trusting less credible information. A promising direction to address this problem is the development of support tools for information credibility. Such tools already have been researched and developed, for instance, aiming at the credibility of Wikipedia (e.g., Adler et al., 2008; Korsgaard & Jensen, 2009). The relationship between advice given by such support systems and users' own assessments should be examined. The 3S-model can provide factors to consider in such examinations. It is plausible that users who mainly use their source experience benefit more from support systems than do users who actively assess the information themselves. Differences also may be found between domain experts and novices.

4.3 Conclusion

This study has provided new insights concerning the concept of domain expertise in trust judgment behavior of Internet users. A new model of trust judgment has been proposed in which three distinct strategies are identified. Users rely on their domain expertise, information skills, or experience with the source of information to form a trust judgment. An initial validation has been performed, mainly focusing on domain expertise. More empirical studies focusing on other components of the 3S-model are necessary. Knowing these strategies, we more clearly understand how trust judgments are formed. Furthermore, we are more able to predict the information features on which trust judgments depend. The proposed 3S-model can serve as a framework for further research on trust in information and support systems.

Chapter 3

Topic Familiarity and Information Skills in Online Credibility Evaluation

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, namely topic familiarity and information skills. Participants with varying expected levels of information skills (i.e., high school students, undergraduates, and post-graduates) 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 to 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.

Published as Lucassen, T., Muilwijk, R., Noordzij, M. L., & Schraagen, J. M. (2013). *Journal of the American Society for Information Science and Technology*.

1. Introduction

Nowadays, we live in a world in which anyone can go online to attain all information imaginable, and more. While 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 often 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 (Metzger, 2007; Hilligoss & Rieh, 2008), information features

(Yaari, Baruchson-Arbib, & Bar-Ilan, 2011; Lucassen & Schraagen, 2010), 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, 2011b). 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 conducted can show which particular strategies to evaluate credibility are applied by various users.

The remainder of this paper 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, 2011b) 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 paper ends with a discussion of the results and their implications for academic research and practice.

1.1 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).

In the case of 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, 2011b). 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 for 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. 2087). 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.

1.2 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 (2011b). In this model, 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 attended: 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). However, the strategy of considering surface features requires different skills from the user, namely, generic information skills. Such skills include knowledge of the user on 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 unchanged, but two minor adjustments have been made to enhance the clarity of the visualization.

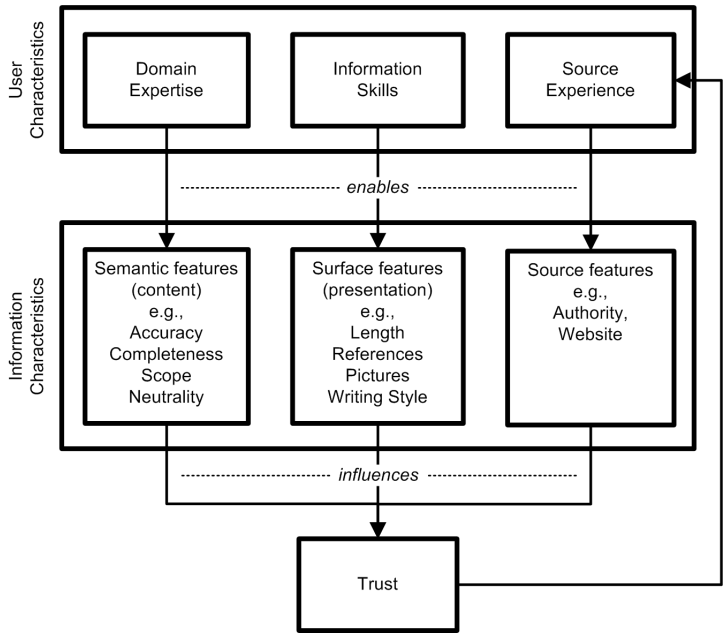

Figure 1: Revised 3S-model of credibility evaluation

First, the originally introduced 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). Since 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, but only where it came from. Earlier experiences with this source are in this case a predictor of the credibility of the current information.

1.3 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 hand leads to more trust (Self, 1996; Chesney, 2006) in the information. However, we argue that this is not necessarily the case. Following Lucassen and Schraagen (2011b), 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.

1.4 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 of experts.

As follows from 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 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 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).

1.5 Wikipedia

One particularly interesting source on the web to study the evaluation behavior of lay internet 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., Giles, 2005; Chesney, 2006; 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 imposes 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 on this source. Lucassen and Schraagen (2011b) 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 high- and low-quality information on Wikipedia (Lucassen & Schraagen, 2010).

1.6 Hypotheses

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

> *H1: 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 aids 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:

> *H2: 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.

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." While 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:

H3: Trust of users with better information skills is influenced by the quality of the information.

H4: 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.

2. Method

2.1 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 six 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 2–5 years with an average of 4. Only three of the high school students mentioned the open-editing model behindWikipedia when asked to explain the basics of this website. One high school student had experience in editing articles on Wikipedia himself.

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

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.

The undergraduates were all following education in the domain of behavioral sciences. They received course credits for participating. Their experience with Wikipedia ranged from 3–8 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 toWikipedia before.

The postgraduates were from various disciplines, such as behavioral sciences, physics, and management sciences. Their experience with Wikipedia ranged from 4–10 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.

2.2 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.

2.3 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.

Article quality was manipulated following the classification of the Wikipedia Editorial Team ("Wikipedia:Version 1.0 Editorial Team," n.d.). 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 highest three 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.

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.

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

Unfortunately, no classification of quality is available on the DutchWikipedia (apart from the Dutch equivalent of the "Featured articles," but these are very few). Instead, we applied the clear criteria of theWikipedia 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 doublerating 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 onWikipedia, with the exception of the removal of cues specific for Wikipedia, indicating diminished credibility (e.g., [citation needed] indications) or high credibility (e.g., bronze stars in Featured articles). The 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.

2.4 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 under 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 withWikipedia (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.

2.5 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, 2011b) 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 doublecoded 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 greatly differed 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:

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

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

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).

3. Results

3.1 Familiarity Manipulation Check

The questionnaires after each article indicated that the manipulation of familiarity was successful. On a 1–7 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).

3.2 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 = 1122) = 41.74, $p < .001$; postgraduates: χ^2 (1, N = 1010) = 29.43, $p < .001$).

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

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

Percentages are given in parentheses.

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 = 2793) = 435.85, 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.

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

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

Percentages are given in parentheses.

Table 5: Key features used by participants of 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	

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

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

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

3.3 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) = 0.21, 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).

4. Discussion

In this study the influence of domain expertise and information skills on credibility evaluation and trust was 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 lowquality 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, 2011b), it was already shown that trust of domain experts was influenced when a semantic feature (enabled for domain experts) indicated low credibility. Now, we also demonstrated that when surface features indicate lower credibility, this only has an influence on users with 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 to a certain user. Only when a cue is prominent can the user give an interpretation to this cue (i.e., consequences for credibility), and have an influence on 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 on the topic at hand 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 (2011b), 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 than 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 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 on 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 as 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 (Self, 1996; Eastin, 2001; Chesney, 2006), 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 1–7 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, since 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 withWikipedia may increase trust in this source (Lucassen & Schraagen,

2011b). 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).

4.1 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 of 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.

4.2 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, 2011b). 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. Also, we thank C.S.G. Dingstede for allowing us to perform the experiments with high school students at their school.

Chapter 4

Reference Blindness: The Influence of References on Trust in Wikipedia

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.

Published as Lucassen, T., Noordzij, M. L., & Schraagen, J. M. (2011). Reference Blindness: The Influence of References on Trust in Wikipedia. In *ACM WebSci '11, June 14-17 2011*, Koblenz, Germany.

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 (Waters, 2007; Dooley, 2010). Nevertheless, the quality of the articles has been proven to be quite high in comparison to professionally maintained databases (Rajagopalan et al., 2010). It has even been shown that the accuracy of Wikipedia is similar to a traditional encyclopedia (Giles, 2005). However, due to the open editing model, the risk of encountering false information is always present (Denning, Horning, Parnas, & Weinstein, 2005). 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 (Fogg & Tseng, 1999). 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 (Fogg, 2003); examples are text length, writing style, or images (Lucassen & Schraagen, 2010).

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 (Chaiken & Maheswaran, 1994) can be helpful. It has been proposed that credibility evaluation can be carried out in a peripheral/heuristic or central/systematic manner (Fogg & Tseng, 1999; Metzger, 2007; Hilligoss & Rieh, 2008). 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.

It has been shown that college students are capable of performing meaningful credibility evaluations of Wikipedia articles (Lucassen & Schraagen, 2010). 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

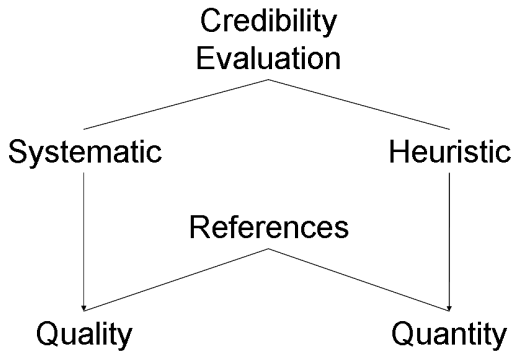

Figure 1: The relationship between systematic and heuristic evaluation and the corresponding features of references.

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 (Lucassen & Schraagen, 2010), this leads to the following two hypotheses:

H1: Reference quality has a positive impact on information credibility when a systematic evaluation is performed.

H2: 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 (Lucassen & Schraagen, 2011b). 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:

H3: Users with high trust in the source perform a heuristic credibility evaluation.

H4: 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 (Metzger, 2007). 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 in 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 (Lucassen & Schraagen, 2010).

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 (Lucassen & Schraagen, 2010). In this task, a Wikipedia article is presented, in which

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

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 (Wikipedia: Version 1.0 editorial team, n.d.) 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.

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 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.

Articles with references of high quality were trusted more by the participants than articles with low-quality references ($t(22) = 3.07$, $p < .01$), indicating that systematic evaluations were performed during the experiment. Articles with long reference lists were trusted more than articles with short reference lists ($t(22) = 2.05$, $p < .05$), 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 ($t(9) = 2.85$, $p < .05$), whereas no effect of the quantity manipulation was found ($t(9) = 1.00$, $p = .17$). 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 ($t(9) = 1.80$, $p < .05$). 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.

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)

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 five 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 (Lim, 2009), 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.

Acknowledgments

The authors would like to thank Merel Jung and Rienco Muilwijk for their efforts in gathering data.

Chapter 5

The Influence of Source Cues and Topic Familiarity on Credibility Evaluation

An important cue in the evaluation of the credibility of online information is the source from which the information comes. Earlier, it has been hypothesized that the source of information is less important when one is familiar with the topic at hand. However, no conclusive results were found to confirm this hypothesis. In this study, we re-examine the relationship between the source of information and topic familiarity. In an experiment with Wikipedia articles with and without the standard Wikipedia layout, we showed that, contrary to our expectations, familiar users have less trust in the information when they know it comes from Wikipedia than when they do not know its source. For unfamiliar users, no differences were found. Moreover, source cues only influenced trust when the credibility of the information itself was ambiguous. These results are interpreted in the 3S-model of information trust (Lucassen & Schraagen, 2011b).

Submitted as Lucassen, T. & Schraagen, J. M. (n. d.). *Computers in Human Behavior.*

1. Introduction

With the rise of Web 2.0 and user-generated content, the need for a solid evaluation of credibility of information is also becoming increasingly urgent. This is potentially problematic, as users are often not trained for this task.

A traditional heuristic of evaluating the credibility of information is to consider its source. When a source has proven to be credible in the past, it is often assumed that it will be in the future. However, for sources (e.g., websites) with collaboratively generated content, this may not always be the case, as the authors of the content are often unknown. Moreover, different authors may contribute over time. Still, the diminished predictability of this heuristic does not mean that it is not used anymore (Lucassen & Schraagen, 2011b).

In an earlier study, it was hypothesized that (pre-existing) knowledge of the content of the information that is searched would interact with the source's influence on credibility (Eastin, 2001). That is, when someone has more knowledge of the content, the credibility of the source of this information is less important for the overall credibility of the information, as ´expert´ users can judge the credibility of the information themselves. However, Eastin (2001) did not find support for the hypothesized interaction.

In this study, we reinvestigate whether the presence of source cues (e.g., layout and logo of a website), influences the credibility evaluation of online information. Since the original study in 2001, the online environment has changed considerably. One particularly important development has been the introduction of Web 2.0 (Cormode & Krishnamurthy, 2008), with which many websites have started to feature user-generated content. We take one particular website with such an open-editing model , namely Wikipedia, as a case study. In an experiment, information on familiar and unfamiliar topics with or without the layout (source cues) of Wikipedia was offered to participants. Participants were asked to evaluate the credibility of the information presented.

The rest of this paper is structured as follows. We start with a general discussion on credibility evaluation. Next, we discuss the roles of the source of information and domain expertise on the topic at hand, and what can be expected in the context of Wikipedia. Our hypotheses are stated next, followed by our methodology to examine them. After this, we present our results and reflect on our findings in the discussion.

1.1 Credibility Evaluation

How lay people cope with the task of evaluating credibility has been the topic of research for some time now. As a result, several models have been proposed which explain and predict the behavior of people putting trust in information found online.

First of all, it is important to define the concept of credibility. Fogg and Tseng (1999) describe credibility as the *believability* of information. It is generally accepted as the *perceived quality* of information. Credibility of information is based on two key aspects of its author, namely *trustworthiness* and *expertise* (Hovland, Janis, & Kelley, 1982). When evaluating credibility, people consider various cues of the information (or its author) which they deem relevant to the quality of the information. The result of such an evaluation is a level of *trust*. A common definition of trust is given by Mayer, Davis, and Schoorman (1995):

> *The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (p. 712)*

Trust is in turn a decisive factor in the decision to use information or not (Kelton, Fleischmann, & Wallace, 2008). By using information, the risk is taken that the information turns out to be inaccurate. Credibility evaluation aims at reducing this risk.

In an attempt to shed more light on how people perform credibility evaluations, the 3S-model of information trust was first introduced by Lucassen and Schraagen (2011b) and later refined by Lucassen, Muilwijk, Noordzij, and Schraagen (2013). Three strategies were distinguished that users apply to evaluate credibility of information, two active (involving actual evaluation) and one passive (relying on earlier experiences). First, the most intuitive active approach is to evaluate the semantics of the information. By doing so, the user tries to verify the factual accuracy of the information. However, this requires the user to have a level of domain expertise on the topic at hand. The second active strategy can also be applied when little or no domain expertise is available, namely the evaluation of surface features. Following this strategy, the way in which the information is presented is considered as an indirect cue of credibility.

A third, passive, strategy was also distinguished, namely considering the source of the information. Instead of actively considering the actual content of the information, users

may rely on their earlier experiences with the source. This strategy may be considered *passive*, as no actual task is performed to evaluate credibility. Lucassen and Schraagen (2011b) argued that the strategy of considering the source of information is biasing influence on a user's trust judgment.

1.2 Source Credibility

The concept of source credibility has been studied for a long time. Hovland and Weiss (1951) already showed that information from highly credible sources was trusted more than information from sources with low credibility. However, this difference faded with time; information from both source types was trusted equally after a few weeks.

Considering source credibility can be considered heuristic processing in credibility evaluation (Chaiken & Maheswaran, 1994). An important determinant for this behavior is task importance (i.e., the necessity for correct information). When task importance is low, users often perform a heuristic evaluation, thus focusing on the source of information. A systematic evaluation will only be performed when task importance is higher.

Eastin (2001) investigated the relationship between source expertise and knowledge of the content. Source expertise can be considered a more specific form of source credibility, since the key components of credibility are trustworthiness and expertise (Hovland, Janis, & Kelley, 1982; Fogg & Tseng, 1999). An interaction between source expertise and knowledge of content was hypothesized; Eastin (2001) expected that users with more knowledge would attribute less value to source credibility, as they are able to focus more on the actual content of the information. However, no significant interaction was found in an online experiment. The lack of differences found between sources of varying expertise was attributed to the overall high level of trust in online information.

1.3 Domain Expertise

As stated, Eastin (2011) did not find an effect of knowledge of content (or level of domain expertise) on the use of source credibility in credibility evaluation. However, domain expertise has been demonstrated to influence trust in information in various ways.

Lucassen and Schraagen (2011b) showed that domain experts are influenced by arguably the most important aspect of credibility, namely factual accuracy. In contrast, domain novices do not notice factual errors in information. Hence, it could be argued that domain

experts are better calibrated to the accuracy of information. This notion is supported by Kelton et al. (2008), who argued that the general propensity to trust is less influential on trust in a particular piece of information when the user has knowledge on the topic at hand. Moreover, Fogg and Tseng (1999) showed that users of automation evaluate computer products more stringently when they are familiar with the content. As a result, perceived credibility will be lower with familiar users.

Chesney (2006) asked several domain experts to rate the credibility of Wikipedia articles on topics inside and outside their own domain. Information in their own domain was perceived more credible (albeit only at the 10% significance level) as information outside their domain. Based on this difference, Chesney concluded that Wikipedia is highly accurate.

1.4 Wikipedia

Now that the potential influence of the source of information on perceived credibility is established, the influence of Wikipedia as an information source should be considered. This has yielded ambiguous results. In an online experiment by Lucassen and Schraagen (2011b), about 30 percent of the participants noted the source of the information (Wikipedia) as a motivation for their level of trust. However, for about half of these participants, the source was a reason to trust the information, whereas the other half did not trust the information for the same reason.

Kubiszewski, Noordewier, and Costanza (2011) systematically manipulated the source of information by presenting the same information in the form of a page on Encyclopædia Britannica, Wikipedia, or Encyclopedia of Earth to their participants. They found that information presented on Encyclopædia Britannica was perceived as having significantly more credibility than on the other two sources. A similar effect was found by Flanagin and Metzger (2011). However, they also demonstrated that while information presented *on* Wikipedia was perceived as less credible (regardless of the content), information actually *from* Wikipedia was perceived as more credible (regardless of the source presented). This clearly demonstrates the distinction between the active and passive strategies in the 3S-model by Lucassen and Schraagen (2011b); when actively evaluating the information itself the perceived credibility is different than when passively relying on earlier experiences with the source of information.

1.5 Hypotheses

It is thus hard to predict whether the availability of source cues indicating that information comes from Wikipedia actually *increases* or *diminishes* the credibility of the information. However, compared to a condition without any cues about the source of the presented information, we expect that trust will be higher. When the source of information is known, users have more cues to base their trust on (cf. 3S-model, Lucassen & Schraagen, 2011b), which is likely to make them more confident of the credibility of information. This leads to the first hypothesis:

H1: Information is trusted more when the source (Wikipedia) is known.

In line with the hypothesized interaction between source credibility and knowledge of content (Eastin, 2001), we expect that the availability of source cues is less important for people who are familiar with the topic at hand. Such familiar users are able to directly evaluate semantic features of the information (Lucassen & Schraagen, 2011b), which makes them less dependent on source cues. This leads to the second hypothesis:

H2: Users who are familiar with the topic at hand are less influenced by source cues than unfamiliar users.

A second influence of source cues on credibility evaluation can also be expected. Considering the source of information as an indicator of credibility is highly heuristic behavior (Chaiken & Maheswaran, 1994). As the motivation to evaluate credibility is often limited (Metzger, 2007), the availability of this heuristic may lead users to skip other, more effortful evaluation methods altogether (i.e., evaluating the actual information itself). This possibility was also acknowledged by Lucassen and Schraagen (2011b), who described the 'source heuristic' as a biasing influence on trust. When the availability of source cues indeed leads users away from a more systematic evaluation, the actual quality of the information is likely to have less influence on trust, as the cues which indicate the quality level are not considered. This leads to the final hypothesis:

H3: The influence of information quality on trust is diminished by the availability of source cues.

2. Method

2.1 Participants

A total of 43 university students participated in the experiment for course credit. Two of these students had to be removed from the analysis due to incomplete data. The average age of the remaining 41 students was 20.6 years (SD = 2.1). The participants were either Dutch (N = 29) or German (N = 12). The questionnaires in the experiment were in Dutch, whereas the stimulus articles were obtained from the English Wikipedia. All participants were familiar with reading texts in these two languages. Hence, the participants reported no major language problems after taking part in the experiment.

2.2 Task and Design

In the experiment, the participants were asked to rate several aspects of a piece of information. These aspects were trust, readability, familiarity, information quantity, writing style, and structure. Answers could be provided on 7-point Likert scales. In fact, we were only interested in trust and familiarity. However, these aspects were boxed in by the other questions to avoid overemphasis on trust. Rated familiarity was used as manipulation check for the familiarity manipulation.

A 2 (source cues) × 2 (familiarity) × 2 (quality) design was used. The availability of source cues was a between-subjects factor whereas the other two manipulations were within-subject. The order of familiarity and quality was counter-balanced to control for order effects.

The availability of source cues was manipulated by presenting information from Wikipedia in the layout of Wikipedia or in a standardized layout. In this case, the standard layout of Microsoft Word 2003 was selected. By using the Word layout, the participants were not able to see that the information came from Wikipedia or any other source. All elements of the original articles were included in the Word-versions. Images and tables were presented in line with the text on the same locations and in the same sizes as on Wikipedia. The articles were presented on a 17" CRT-screen in either Microsoft Internet Explorer or Microsoft Word 2003.

The topics of the articles used were selected such that half of them were assumed to be relatively familiar to the participants and the other half relatively unfamiliar. Table 1 shows

Table 1: Selected topics. All articles were obtained from the English Wikipedia on January 1st, 2011.

Topic	Familiarity	Quality
The Simpsons	Familiar	High
Facebook	Familiar	High
Classical Conditioning	Familiar	Low
Birthday	Familiar	Low
The retreat of glaciers since 1850	Unfamiliar	High
Agartala	Unfamiliar	High
Sovereignty	Unfamiliar	Low
Gandingan	Unfamiliar	Low

which articles were selected as familiar and unfamiliar. Familiarity was assumed based on the age (about 20 years), culture (Western Europe), and education (college education in Psychology) of the participants.

Article quality was manipulated following the quality assessment ratings of the Wikipedia Editorial Team (Wikipedia: Version 1.0 editorial team, n.d.). Half of the articles were of high quality (featured articles and good articles), whereas the other half of the articles were of low quality (stub articles and start articles). High and low quality were evenly distributed in the familiar and unfamiliar condition. Table 1 shows the familiar and unfamiliar articles used in the high and low quality conditions.

2.3 Procedure

Participants were welcomed and asked to sign informed consent. Task instructions were given verbally, based on a standard protocol, and in writing. After this, a questionnaire on various demographics was filled out, after which the experiment started. Each participant performed the task eight times, with a questionnaire on the dependent variables presented after each trial. The experiment ended with a few control questions on the manipulations. The duration of the experiment was about sixty minutes.

3. Results

3.1 Familiarity Manipulation Check

The questionnaires after each article confirmed the manipulation of familiarity. On a 7-point Likert scale the participants rated their prior knowledge on familiar topics (M =

4.88; SD = 0.96) higher than their prior knowledge on unfamiliar topics (M = 2.40; SD = 1.15); $t(40) = 13.65$; $p < .001$.

3.2 Trust

Table 2 shows the level of trust in the various conditions of the experiment. A repeated-measures ANOVA with quality and familiarity as within-subject factors and the availability of source cues as between-subjects factor was performed to test the hypothesis. A main effect of quality on trust was found; high-quality articles were trusted more than low-quality articles; $F(1, 39) = 27.66$, $p < .001$. No main effects were found for familiarity, $F(1, 39) = 0.38$, $p = .54$, and the availability of source cues, $F(1, 39) = 1.23$, $p = .28$. Hence, no support was found for hypothesis 1.

A significant interaction between the availability of source cues and familiarity with the topic was found, $F(1, 39) = 5.07$, $p < .05$. Only trust of familiar users was influenced by the availability of source cues; they trusted information less when they knew it came from Wikipedia. However, an additional three-way interaction with quality revealed that this effect was only visible with low-quality information, $F(1, 39) = 5.72$, $p < .05$. A (two-way) interaction between the availability of source cues and familiarity was hypothesized. However, we expected an inverse effect of what the results showed. Therefore, the second hypothesis has to be rejected.

No other significant interaction effects were found. Therefore, our final hypothesis also has to be rejected, as no interaction between the availability of source cues and information quality on trust was found; $F(1, 39) = 0.25$, $p = .62$.

Table 2: Trust on 7-point Likert scales in all conditions. Standard deviations are given in parentheses.

	With source cues			Without source cues		
	HQ	LQ	Total	HQ	LQ	Total
Familiar	5.35	4.13	4.74	5.60	5.02	5.31
	(1.59)	(1.49)	(1.32)	(0.80)	(1.07)	(0.75)
Unfamiliar	4.33	4.48	4.90	5.60	4.45	5.02
	(1.50)	(1.18)	(1.25)	(0.83)	(1.27)	(0.81)
Total	5.34	4.30	4.82	5.60	4.74	5.17
	(1.52)	(1.25)	(1.26)	(0.72)	(1.03)	(0.68)

4. Discussion

In this study, we investigated the influence of the availability of source cues on credibility evaluations by users who are familiar or unfamiliar with the topic at hand. No support for our stated hypotheses could be found in the conducted experiment. First, no main effect of the availability of source cues on trust was found. Second, concerning the role of familiarity, an effect opposite to our hypothesis was found: only familiar users were influenced by the availability of source cues; unfamiliar users were not. A more detailed analysis revealed that familiar users were only influenced by source cues when viewing low-quality articles. Finally, no biasing effect of source cues on the influence of information quality on credibility was found as high quality information could be distinguished from low quality information in both conditions.

A few explanations can be given for the lack of an effect of the availability of source cues on trust of unfamiliar users. Eastin (2001) argued that in his experiment, no effects were found due to the overall high level of trust in online information. The same could be argued in the current study with Wikipedia as an information source, although the overall level of trust seems to be slightly lower than in Eastin's study. A second explanation could be that knowing that information comes from Wikipedia has a polarizing effect on trust. As was demonstrated by Lucassen and Schraagen (2011b), the source (Wikipedia) was used often (about 30 percent of the cases) as a motivation to both trust and distrust information. It could thus be the case that about half of our participants trusted the information less because it came from Wikipedia, whereas the other half trusted the information more. However, although the standard deviation is slightly larger in the ´Wikipedia-condition´, visual inspection of the distribution of trust scores did not reveal such a polarization effect.

A perhaps more plausible explanation may be derived from the reputation of Wikipedia, or the lack thereof. As shown in earlier studies (e.g., Lucassen et al., 2013), college students are well aware of open-editing model behind Wikipedia. This could mean that our participants knew that they could not attribute any value to the authors of the articles they read. If this is the case, then the condition without source cues was in fact very similar to the condition with source cues; no inferences were made about the source of the information. Hence, evaluation behavior and trust of our participants was the same on Wikipedia and in the condition without cues on the source of information. So it could be argued that for our (unfamiliar) participants, Wikipedia has no reputation at all.

The finding that users familiar with the topic at hand, as opposed to unfamiliar users, were influenced by the presence of source cues, is contradictory to our hypothesis. Our expectations were based on a study by Eastin (2001). He hypothesized the same effect as we hypothesized in this study, but also did not find corroborating results. However, individual tests of both groups (i.e., familiar and unfamiliar) indicated some validity of his original hypothesis.

An explanation for the contradictory findings in this experiment can be found in the difference in our conceptualization of source cues compared to Eastin's study. He manipulated source expertise by presenting the same message from persons who could be expected to vary in their level of knowledge on the topic. In the present study, the source was manipulated by taking articles from Wikipedia, and removing any cues of this source to create a condition without source cues. However, the credibility of Wikipedia is ambiguous, with evidence in favor (e.g., Giles, 2005) and against (e.g., Denning, Horning, Parnas, & Weinstein, 2005) its credibility, both likely to be caused by its open-editing model. Hence, the fact that the information came from Wikipedia did not make the information more credible per se. In fact, the open editing model behind Wikipedia may be an explanation for the observation that users familiar with the topic at hand had less trust in the information when they knew it came from Wikipedia. Since only very little is known about the authors of Wikipedia articles (anyone can contribute), familiar users may have doubted the expertise of these authors (one of the two elements of credibility; Fogg & Tseng, 1999; Hovland, Janis, & Kelley, 1982) in comparison with their own knowledge on the topic. Unfamiliar users do not possess such knowledge, which may have led them to believe that the authors of Wikipedia at least know more about the topic than they do themselves. In their case, this may not have raised trust, but at least it was not diminished.

This explanation seems to be at odds with the study by Chesney (2006), who found that experts had more trust in Wikipedia than novices. However, the experts in that study were actual renowned domain experts on the topics of the articles. This means that those participants were actually able to recognize that information was correct or incorrect. In this study, the 'expert' participants were merely familiar with the topic at hand. This means that they could not solely rely on their own expertise and had to consider other cues as well. One of these cues is the presumed credibility of the authors of information.

As can be seen in the results, the interaction effect between familiarity and source cues on trust was only present in low-quality articles. An explanation for this observation can

again be found in the ambiguity of the credibility of the information. The number of cues in the high-quality articles indicating high credibility may have been so overwhelming to our participants, that the lack of information on its source could not diminish their (positive) opinion. In contrast, the credibility of low-quality articles was much more ambiguous. In these articles, cues indicating both high credibility (e.g., no large errors) and low credibility (e.g., short text) could be found. In this case, it is much more likely that the opinion of the participants was also influenced by the source. In other words: source cues are much more influential when the credibility of the information is ambiguous in itself.

No biasing effect of source cues on credibility evaluation was visible in terms of difference in trust in high and low quality information. This means that our participants were always able to distinguish high and low quality, regardless of its source. While it has been shown before that college students (such as our participants) are able to make such distinctions (Lucassen & Schraagen, 2010), it does not necessarily mean that they also make such distinctions in actual search behavior online. For instance Walraven, Brand-Gruwel, and Boshuizen (2009) have shown large differences between reported and actual evaluation behavior of students. Moreover, not all Internet users are expected to have similarly advanced information skills as college students, who receive specific training in this task (Lucassen & Schraagen, 2011b).

The results of this experiment demonstrate an important aspect of the 3S-model (Lucassen & Schraagen, 2011; Lucassen et al., 2013). The influence of the three proposed user characteristics (domain expertise, information skills, and source experience) on credibility evaluation have been shown. However, the current experiment has also shown that these characteristics are not isolated from each other. In fact, specific user characteristics do not only influence which *information* characteristics are incorporated in credibility evaluation, but also have an effect on the *other* user characteristics. In this case, two potential influences can be observed.

First, we argued that the information skills of our participants (college students) may have been the reason that source cues indicating that the information came from Wikipedia did not raise or lower trust. In an earlier online study, such an effect was found (Lucassen & Schraagen, 2011b), but it can be argued that the information skills of these participants (a broad sample of general Internet users) were overall less advanced. This may mean that they were less informed about the risks of using an open-editing encyclopedia. Hence, the

overall higher level of information skills of the participants in the current study influenced their perception of the source at hand (Wikipedia) in the sense that no authority was attributed to this particular source.

Second, we have shown that topic familiarity (a weaker form of domain expertise) influenced the importance of source cues. Familiar users were more skeptical of Wikipedia, presumably because of the assumed knowledge of the authors behind this source. It can thus be argued that users that are familiar with a particular topic place less trust in a source if the source is known to be a collaborative repository rather than completely unknown.

In conclusion, this study provides additional corroboration of the user characteristics as proposed in the 3S-model (Lucassen & Schraagen, 2011b). In addition, we also demonstrate an interdependency of these characteristics. Both domain expertise and information skills influence source experience, and thus the use of source cues in credibility evaluation.

4.1 Future Research

This study has begun to clarify the relationship between source cues, topic familiarity, and trust in online information. Many of the findings are likely to have been caused by the fact that a website featuring user-generated content was used as a case study. Similar experiments with other (authoritative) source could confirm our explanations for these findings. Moreover, future experiments should also focus on other Internet users than college students, since their knowledge on and experience with Wikipedia is expected to be extensive. Other users, who are less familiar with this source may value it differently. Finally, we also suggest to try to design future experiments to be as realistic as possible, in order to ensure external validity. While in this study, we attempted to move some emphasis away from the concept of trust, the imposed task was still quite artificial. Moreover, the participants were very well aware of the fact that they were participating in a lab experiment, which may for instance lead to socially desirable behavior. Online studies for example, could diminish such effects (cf. Lucassen & Schraagen, 2011b).

5. Conclusion

The main contribution of this study is that we have shown that the relationship between topic familiarity and source cues in credibility evaluation is not as straightforward as has been suggested previously. In the case of collaborative, open editing, websites such as Wikipedia, the presence of source cues can have a negative effect on trust of familiar users,

whereas in general, a positive relationship between source cues and familiarity may be expected. Future research should aim at further clarification of the relationship between these concepts. We have also demonstrated that knowledge of the source of information has a much larger influence on trust when the credibility of the information in itself is ambiguous.

Acknowledgments

The authors wish to thank Jerfy ter Bekke, Esther Brink, Marlyn Bijvank, and Niels van de Sande for their help in gathering data.

Chapter 6

Propensity to Trust and the Influence of Source and Medium Cues in Credibility Evaluation

Credibility evaluation has become a daily task in the current world of online information that varies in quality. The way this task is performed has been a topic of research for some time now. In this study, we aim to extend this research by proposing an integrated layer model of trust. According to this model, trust in information is influenced by trust in its source. Moreover, source trust is influenced by trust in the medium, which in turn is influenced by a more general propensity to trust. We provide an initial validation of the proposed model by means of an online quasi-experiment (n=152) in which participants rated the credibility of Wikipedia articles. Additionally, the results suggest that the participants were more likely to have too little trust in Wikipedia than too much trust.

Published as Lucassen, T. & Schraagen, J. M. (2012). *Journal of Information Science, 38,* p. 566-577.

1. Introduction

Credibility evaluation in online environments has been shown to be a largely heuristic process (Taraborelli, 2008; Metzger, Flanagin, & Medders, 2010). Internet users are not willing to spend a lot of time and effort on verifying the credibility of online information, which means that various rules-of-thumb are applied to speed up the process. One important strategy is to consider the source of the information (Chaiken & Maheswaran, 1994). In the pre-Internet era, this was a solid predictor of credibility, but nowadays it is hard to point out one single author as being responsible for the credibility of information. Sources are often 'layered' (Sundar & Nass, 2001; Kang, Bae, Zhang, & Sundar, 2011), multiple authors collaborate on one piece of information, and with the advent of Web 2.0, it is often unclear who actually wrote the information.

The diminished predictive power of the credibility of a source could mean that people no longer use it. However, research on online credibility evaluation has shown otherwise. Consider, for instance, the case of Wikipedia. It was shown that numerous Internet users made their decision to trust (or not trust) articles from this source solely based on the fact that they came from Wikipedia (Lucassen & Schraagen, 2011b). For trusting users, considering the source means that they are also likely to trust the occasional poor-quality information from this source (i.e. potential overtrust). In contrast, distrusting users miss out on a lot of high-quality information (i.e. potential undertrust). Hence, the diminished predictive power of the source does not mean that it is no longer used.

Trust in multiple, comparable sources may generalize to trust in a medium (Dutton & Shepherd, 2006). An example of such a medium is the Internet as a generalization of several websites. It has been shown that people often refer to 'the Internet' or even 'the computer' as the source of information they found online, rather than a specific website Sundar (2008). This generalization may be the reason why users have already established a baseline of trust when encountering new sources of the same type (i.e. websites).

In this study, we examine the influence of trust in the source and trust in the medium on credibility evaluation. A more general propensity to trust is also considered, as this may serve as a disposition for more case-specific trust (i.e. trust in a medium, source or piece of information). We hypothesize a layer model in which each type of trust influences the next (see Figure 1). The core of this model is trust in a particular piece of information, which is influenced by trust in the source of this information (Chaiken & Maheswaran,

1994; Lucassen & Schraagen, 2011b). Trust in the source is seen as a specification of trust in a medium (as a collection of sources). Therefore, trust in the medium may serve as a baseline for trust in a source. Furthermore, we hypothesize that trust in a medium is influenced by a user's propensity to trust. Overall, we theorize that trust becomes more specified with each layer; each preceding layer serves as a baseline for the subsequent layer. The proposed model can help us better understand how trust in information is formed.

We study these influences through an online quasi-experiment in the context of the Internet (as a medium) and Wikipedia (as a source), starting with a discussion on each proposed layer of trust individually, after which we present our research model in which we combine them. We introduce three hypotheses, aimed at validating the research model. Next, we describe our methodology to test the hypotheses, followed by the results. Finally, the results are discussed, limitations are identified, and conclusions are drawn.

1.1 Trust in Online Environments

A common definition of trust is 'the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party' (Mayer, Davis, & Schoorman, 1995). This definition implies that a certain risk is taken when someone trusts someone else (Kelton, Fleischmann, & Wallace, 2008); this trust may prove to be unjustified. In the online domain, trust is an especially relevant concept, as Internet users often interact with parties they do not have prior experiences with. Consider, for instance, online financial transactions (e.g. buying a product through a web shop): the consumer is at risk of losing their money when the vendor fails to meet their expectations.

Four levels have been proposed at which trust may be studied (Kelton, Fleischmann, & Wallace, 2008), namely individual (as a personality trait), interpersonal (one actor trusting another), relational (mutual trust) and societal (trust in a community). When considering trust in information, the appropriate level is interpersonal trust, as the reader puts their trust in the author of the information.

In order to reduce the risk associated with trusting someone or something, a credibility evaluation may be performed. In such an evaluation, the 'trustor' searches for cues on the credibility of the 'trustee'. Such evaluations are largely heuristic processes, as the user often lacks the motivation and/or ability for a systematic (thorough) evaluation (Metzger,

2007). According to the MAIN model (Sundar, 2008), today's information technology has resulted in numerous affordances in which credibility cues can be found. Such cues may trigger cognitive heuristics; simple judgment rules to estimate the various dimensions of the quality of information. These dimensions also play an important role in the judgment of credibility.

1.2 Trust in Information

Another model that clarifies the use of various cues of credibility is the 3S-model of information trust (Lucassen & Schraagen, 2011b). This model asserts that the most direct strategy for evaluating credibility is to search for semantic cues in the information. By doing so, Internet users try to answer the question: 'Is this information correct?' Cues such as factual accuracy, neutrality or completeness of the information are considered by users who follow this strategy. This implies that some domain knowledge of the topic at hand is required. However, a typical information search concerns information that is new to the user, as it normally does not make any sense to search for information one already has. This means that users may often lack the required domain expertise to evaluate the semantics of the information, which makes it impossible to apply this strategy.

To work around this deficit, users may also consider surface cues of the information to evaluate credibility. This strategy concerns the manner of presentation of the information. Examples of cues evaluated when following this strategy are the writing style, text length or number of references in the information. While it is a less direct way to evaluate credibility, no domain knowledge is needed. Instead, by considering surface cues, users bring to bear their information skills. Such skills involve knowledge of how certain cues (e.g. a lengthy text, numerous images) relate to the concept of credibility.

Following dual-processing theory (Chaiken, 1980), it is tempting to see the strategy of evaluating semantic cues as a systematic evaluation and the strategy of evaluating surface cues as heuristic. However, both strategies can be performed at various levels of processing. For instance, recognizing something one already knows (semantic) is considered largely heuristic behaviour (Klein, Calderwood, & Clinton-Cirocco, 1986). On the other hand, checking the validity of each of the references of an article on Wikipedia (surface) can be seen as largely systematic. The choice between systematic or heuristic processing in credibility evaluation primarily depends on the motivation and ability of the user (Metzger, 2007).

Thus, the cues used in credibility evaluation depend heavily on user characteristics. This has also been proposed in the unifying framework of credibility assessment (Hilligoss & Rieh, 2008), which suggests that there are three levels of credibility evaluation between an information seeker and an information object (namely the construct, heuristics and interaction layers). The first layer is the construct layer, in which it is posited that each user has their own definition of credibility, which means that different elements of the information object are salient to different users when evaluating credibility (see also Fogg, 2003).

1.3 Trust in the Source

A third, more passive way of evaluating credibility is also posited in the 3S-model (Lucassen & Schraagen, 2011b), namely the strategy of considering the source of information. Following this strategy, earlier interactions with a particular source may serve as a cue for the credibility of the current information. For instance, if someone has numerous positive experiences with information from a particular website, this user may choose to trust new information from that source without actively evaluating its credibility. The opposite is also possible: when one has negative experiences with a source, one may choose to avoid new information from this source without even looking at it (at the semantic or surface level).

The approach of transferring the credibility of the source of information to the credibility of the information itself only works well when the credibility of information from a source is stable over time. However, in online environments, information from one source may vary greatly in credibility. Consider again Wikipedia: information quality is generally very high (Giles, 2005), but numerous examples of incorrect information from Wikipedia are readily available (Cross, 2006; Waters, 2007; Dooley, 2010). This means that trusting this source involves taking the risk of encountering false information. On the other hand, distrusting this source means that the user may miss out on much high-quality, valuable information. Nevertheless, it has been shown that, also in the case of Wikipedia, the source-strategy is applied very often. It was found that around 25% of experts on the topic and 33% of novices trusted or distrusted information solely because it came from Wikipedia (Lucassen & Schraagen, 2011b). This is an indication that users weighed the benefits of Wikipedia (much information) against its risks (poor information). For some users, the benefits clearly outweighed the risks. For others, they did not.

A second drawback of considering the source of information is that nowadays it is often difficult to determine a single author who is responsible for the information. Online news, for instance, is often carried through multiple sources (e.g. a blogger writing a piece on something she read on Facebook, which was a reaction on an article on the CNN news page). This concept is known as 'source layering' (Sundar & Nass, 2001), and makes it increasingly difficult to determine which source is responsible for the credibility of the information. However, a recent study on this phenomenon (Kang et al., 2011) has shown that only highly involved users considered more distal sources when evaluating credibility. Users with low involvement were only influenced by the credibility of the most proximate source (i.e. the website on which they read the news). For this reason, we consider the most proximate source (i.e. Wikipedia) as 'the source' of information in this study.

Two key factors for the credibility of a source have been identified (Fogg & Tseng, 1999). First, sources should have the appropriate knowledge (expertise) to provide correct information. For instance, a doctor is able provide credible health information, whereas a patient may not be. Second, sources should be trustworthy, that is, have the intention to supply correct information. To clarify this concept, consider the difference between a manufacturer of a product and an independent party testing this product. Both may provide similar information about the product, but have very different intentions. The manufacturer wants to sell the product, whereas the tester wants to provide consumer advice. This may have large consequences for the credibility of the information supplied.

1.4 Trust in the Medium

Traditional linear communication models generally encompass a source (sender) of information, who transmits a message through a medium to a receiver. However, it has been shown that a medium may also be treated as a more general type of source by information seekers (Sundar & Nass, 2001). People tend to say that they got information 'off the Internet', or even 'off the computer' rather than naming one specific website (Sundar, 2008). As such, credibility may also be attributed to a medium rather than a single source.

Examples of different media channels are the Internet (or a subset, such as Internet vendors; Lee & Turban, 2001), television, newspapers or school books. It has been shown that trust in the Internet is primarily influenced by experience (Dutton & Shepherd, 2006). It is hardly possible to assign a value to the credibility of online information without having used the Web. Such experience with the Web means that users have interacted with

various online sources (websites). The experiences in these interactions are accumulated into trust in the Internet as a whole.

Trust in a medium can be brought to bear when encountering a new source on this medium (i.e. an unfamiliar website). Users may evaluate the credibility of this website, as well as the information on it, but trust in the Internet in general may serve as a baseline.

In the context of research on the Internet, this medium is often compared with traditional sources such as books or newspapers (Wathen & Burkell, 2002). Differences are found at various levels, such as organization, usability, presentation and vividness. In various instances, the Internet has been shown to be more credible (e.g. political information; Johnson & Kaye, 1998) and less credible (e.g. health information; McClung, Murray, & Heitlinger, 1998) than traditional media.

1.5 Propensity to Trust

As stated earlier, when studying trust in information, the appropriate level is interpersonal trust (Kelton, Fleischmann, & Wallace, 2008). However, this does not mean that trust on the other levels has no influence. Consider, for instance, trust at the individual level, or 'propensity to trust'. Propensity to trust is a personality trait, a stable factor within a person, that affects someone's likelihood to trust (Lucassen & Schraagen, 2011b).

One's propensity to trust, or dispositional trust, serves as a starting point, upon which more case-specific trust builds (Merritt & Ilgen, 2008). In an experiment with an X-ray screening task with automation available, trust in the automation moved from dispositional to history-based (Merritt & Ilgen, 2008). In other words, trust became more calibrated to the automation. Owing to the heuristic character of online credibility evaluation, we expect that the propensity to trust will also be visible in more case-specific trust, such as trust in the medium, source or actual information.

The relationship between propensity to trust and trust in online information has been studied before. Propensity to trust has been shown to be among the most influential factors predicting consumers' trust in Internet shopping (Lee & Turban, 2001; Cheung & Lee, 2001). In these studies, propensity to trust is seen as a mediating factor between trustworthiness of the vendors and the external environment on the one hand, and trust on the other. Some researchers distinguish a propensity to trust from a propensity to distrust (McKnight, Kacmar, & Choudhury, 2004). Again, in the context of e-commerce,

it was shown that the former has an influence on trust in low-risk situations, whereas the latter influences trust in high-risk situations.

It is not surprising that much research on trust in online environments has focused on e-commerce. In this domain, users take a direct, measurable risk (of losing money), which makes trust a very important construct. This risk may be less salient (or at least measurable) in other domains, such as online information search, as it heavily depends on the purpose of the information. However, wrong decisions may be taken based on this information, which makes online information search an important area of study.

1.6 Proposed Research Model

In the literature discussed here, the concepts of trust in information, sources and media, and a general propensity to trust are mostly studied in isolation from each other. Some exceptions can be found (e.g. Lucassen & Schraagen, 2011b; Sundar, Knoblock-Westerwick, & Hastall, 2007), but an integrated approach featuring all these concepts in one study is yet to be seen. In this study, we present a novel model of trust in information, explaining how these concepts are related to each other.

As presented in Figure 1, we hypothesize that not all types of trust discussed here have a direct influence on trust in each particular piece of information. Instead, we suggest a layer model, in which trust is built from a general propensity to trust to case-specific trust in a particular piece of information. In this model, we consider general propensity to trust as the general baseline of trust of a person (Merritt & Ilgen, 2008) in all situations, hence not only for trust in (online) information, but also trust in, for example, others, society or technology. With each layer, trust becomes more specific for a single situation (i.e. evaluating the credibility of a single piece of information).

The second layer is labelled 'trust in the medium' and concerns trust of a user in a particular type of medium (e.g. newspapers, the Internet). While this is clearly a more case-specific form of trust than a general propensity (at least one feature of the information at hand is considered), it is still a generalization of trust in the source of the information (Sundar & Nass, 2001). Trust in a medium can also be seen as trust in a collection of sources.

Trust in the medium is followed by the layer 'trust in the source'. Again, trust is further specified, as the specific source of the information is considered rather than the medium through which the information is communicated. Considering the source of information

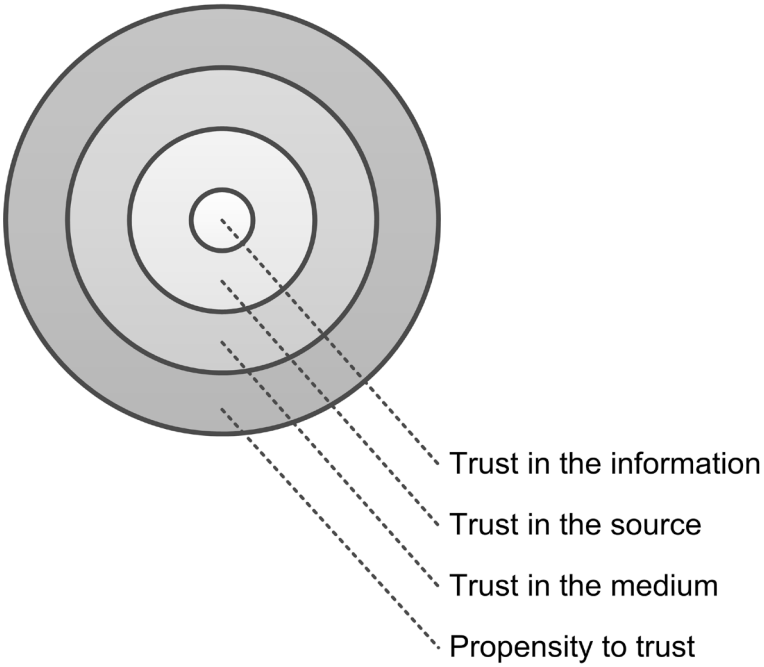

Figure 1: Proposed 'layer' model of trust. Each layer is a further specification of trust, and influences the next layer.

is perhaps the most traditional form of credibility evaluation (Lucassen & Schraagen, 2011b).

The most specific form of trust is trust in the information itself. Especially when the credibility of the source is doubtful, users may search for cues in the information itself to estimate credibility (Sundar, Knoblock-Westerwick, & Hastall, 2007). In sources where the credibility varies between different pieces of information (e.g. Wikipedia), trust is best calibrated with the actual information quality when cues from the information itself are considered, rather than cues from the source or medium (Lucassen & Schraagen, 2011b).

In this study, we seek initial validation for this model of trust, by evaluating the influence of each of these layers on trust in Wikipedia articles. Thereby, we assume that all participants actively evaluate the credibility of the information to a certain extent (as this is the task imposed on them). However, the layer at which credibility is evaluated in practice may largely vary between users and contexts. Motivation and ability to evaluate have been identified as important factors for the extent to which credibility is evaluated (Metzger, 2007). Only relying on one's propensity to trust does not require any effort when encountering a piece of information. Each next layer requires more effort from the user to evaluate credibility. Hence, in situations with a low risk of poor information, or with users

with a low motivation or ability, the outer layers may have a larger influence on trust in the information than when the risk, motivation or ability are higher.

We thus hypothesize a direct influence of each layer on the next. Moreover, we expect that the influence of each layer can also be observed in more distant layers (e.g. the influence of trust in a medium on trust in information). However, we hypothesize that this influence is mediated by the intermediate layer, which can better explain how the two non-adjacent layers are related (e.g. trust in a source explains the relationship between trust in a medium and trust in information). Hence, the following hypotheses are tested through mediation analysis in order to examine the validity of the proposed research model. The following three hypotheses can be derived from the hypothesized model:

H1: Propensity to trust has a direct positive influence on trust in the medium. Its influence on trust in the source is mediated by trust in the medium; its influence on trust in information is mediated by both trust in the medium and trust in the source.

H2: Trust in the medium has a direct positive influence on trust in the source. Its influence on trust in information is mediated by trust in the source.

H3: Trust in the source has a positive influence on trust in a particular piece of information from this source.

2. Method

2.1 Participants

Invitations for participation in an online experiment were posted on several online forums and social media, and via direct email contact. This resulted in a total of 152 participants who completed the whole experiment. Three participants were excluded for bogus participation; they gave the same answer to every question. Of the remaining 149 participants, the majority (81.9%, n = 122) was male. The average age was 25.7 years (SD = 10.1). The participants came from Europe (67.8%), North America (25.5%), Australia (2.7%), South America (2.0%) and Asia (2.0%). It was ensured that each participant could only partake once by registering their IP-addresses and placing a cookie on their computer.

2.2 Task and Procedure

The experiment was conducted using an online questionnaire. When following the link to the questionnaire in the invitations, an explanation of the task was presented first. Participants were informed that they had to evaluate the credibility of two Wikipedia articles, without specifying how to perform this task (Wikipedia Screening Task; Lucassen & Schraagen, 2010). The explanation also stated that, after the evaluation, a few questions about trust and Internet use would be asked. Moreover, the participants were warned that they could not use the back and forward buttons of their browser.

After reading the instructions, the participants could decide to participate by clicking on the 'next' button. After doing this, they were asked to enter their gender, age, and nationality on the subsequent page.

The actual experiment started after the participants clicked on the 'next' button again. A full-page screenshot of a Wikipedia article was shown in the questionnaire. Underneath the article, the participants had to answer three questions. First, they had to rate how much trust they had in the article on a seven-point Likert scale. Second, they could provide an explanation for their answer through an open-ended question. This explanation was mainly used as an indicator for bogus participation, but the explanations were subsequently also categorized according to the 3S-model. Third, the participants had to rate how much they already knew about the topic at hand, again on a seven-point Likert scale. After answering these questions and clicking 'next', the procedure was repeated for the second article.

After evaluating both articles, three separate webpages asked for (1) propensity to trust in general, (2) trust in Wikipedia and the Internet, and (3) general remarks on the experiment.

2.3 Stimuli

Each participant viewed one article of high quality and one article of low quality. The ratings of the Wikipedia Editorial Team (Wikipedia: Version 1.0 editorial team, n.d.) were used to make this distinction. For the high-quality articles, the highest quality class (Featured Articles) was used whereas for the low-quality articles, the second lowest quality class (Start-class Articles) was used. The lowest quality class (Stub articles) was deliberately avoided, as these are often single-sentence articles.

Next to quality, length of the articles was also taken into account. Some featured articles tend to be extremely lengthy. This could cause problems in the experiment, as it could take too much time for the participants to evaluate such articles. Therefore, only featured articles with fewer than 2000 words were selected. Moreover, we enssured that the articles in the poor-quality condition were sufficiently long to perform a meaningful credibility evaluation (i.e. they contained enough cues to evaluate). Therefore, only start-class articles with more than 300 words were selected.

Following these considerations, three topics with a typically encyclopedic character were used, namely:

- food ('Andouilette' and 'Thomcord');
- historical persons ('Princess Amelia of Great Britain' and 'Wihtred of Kent');
- animals ('Bobbit worm' and 'Australian Tree Frog').

The first of each pair served as a low-quality article, and the latter as a high-quality article. Each participant was randomly assigned to one of the topics and evaluated two articles on this topic. The order of articles was counterbalanced between subjects.

2.4 Measures

2.4.1 Propensity to Trust

Propensity to trust was measured using a subsection of the NEO-PI-R personality test (Costa & McCrea, 1992) regarding trust. This consisted of eight questions, to be answered on five-point Likert scales (see Appendix). Although the NEO-PI-R is not intended for partial usage, we decided not to use the full questionnaire, as this would extend the duration of the experiment substantially, which is not desirable in online experiments. Moreover, the other personality characteristics of the full test did not bear relevance to the scope of this study. A reliability analysis (see Results) ensured the reliability of the remaining questions.

2.4.2 Trust in the Internet

Trust in the Internet was measured on seven-point Likert scales using six questions about (1) usage, (2) perceived credibility, (3) trust in the institutes behind the Internet, (4) confidence in other Internet users, (5) usefulness and (6) privacy protection. Question 2–4, and 6 are based on the Net-confidence and Net-risks scales (Dutton & Shepherd,

2006), extended with questions about usage (1) and usefulness (5), which have been shown to be other salient indicators of trust (Davis, Bagozzi, & Warshaw, 1989; Kelton, Fleischmann, & Wallace, 2008). See Appendix for the full questionnaire.

2.4.3 Trust in Wikipedia

Trust in Wikipedia was measured on seven-point Likert scales using basically the same six questions as used for trust in the Internet, replacing 'the Internet' with 'Wikipedia'. However, some issues could not be easily converted, such as the issue of privacy. Therefore, the nearest related concept applicable to Wikipedia was used (in this case: accuracy). See Appendix for the full questionnaire.

2.4.4 Trust in Information

Trust in the information was measured on a seven-point Likert scale after each article, asking the question 'How much trust do you have in this article?' As each participant viewed one article of high quality and one article of low quality, the average rating was taken for the construct 'trust in information' in the analyses.

2.5 Data Analyses

For each of the constructs measured through questionnaires, its reliability was calculated using Cronbach's α. We took an α of at least .70 as an acceptable value for all three constructs (propensity to trust, trust in the Internet and trust in Wikipedia).

In order to find validation for our research model, bootstrapping mediation analysis was performed (Preacher & Hayes, 2008; Hayes, 2009) to estimate direct and indirect effects with multiple mediators using the PROCESS toolkit for SPSS (Hayes, 2012). The advantages of this technique are that all mediators can be tested simultaneously, normal distribution does not need to be assumed, and the number of inferential tests is minimized (reducing the risk of a type 1 error). Following the proposed research model (see Figure 1), a model with trust in the medium and source as sequential mediators was tested.

The motivations for trust in the articles that could be provided by the participants were classified in accordance with the 3S-model (Lucassen & Schraagen, 2011b). This means that each comment was categorized as referring to a semantic, surface or source feature. Comments that could not be categorized as referring to any of these features were classified as 'other'. Half of the comments were categorized by two raters. Based on this overlap,

Cohen's k was calculated to ensure inter-rater reliability. A *k* of .91 indicated a near-perfect agreement.

3. Results

3.1 Validity of the Questionnaires

Cronbach's α for the participants' propensity to trust derived from the NEO-PI-R questionnaire (Costa & McCrea, 1992) was .82, indicating good reliability. For the trust in the Internet scale, Cronbach's α was .70, indicating acceptable reliability, and for the trust in Wikipedia scale, Cronbach's α was .88, again indicating good reliability.

3.2 Trust

Propensity to trust, as measured through the eight questions on this construct on the NEO-PI-R (Costa & McCrea, 1992), is divided into five categories, displayed in Table 1.

Trust in the Internet ranged from 1.17 to 5.33 (on a Likert-scale from 0 to 6), with an average of 3.63 ($SD = 0.79$). Trust in Wikipedia ranged from 0.00 to 5.83 (on a Likert-scale from 0 to 6), with an average of 3.51 ($SD = 1.11$).

Trust in high-quality information ($M = 4.91$, $SD = 1.64$) was higher than trust in low-quality information ($M = 4.42$, $SD = 1.63$), $t(148) = 2.82$, $p < .01$. Average trust in the information was 4.67 ($SD = 1.24$).

3.3 Validity of the Research Model

Figure 2 shows a cross section of the layer model presented in Figure 1, with unstandardized regression coefficients between all constructs. Trust in the information was entered as the dependent variable, propensity to trust as the predictor variable and trust in the medium

Table 1: Participants in each of the five categories of the NEO-PI-R trust scale.

Category	Percentage
Very high trust	10.7% (n=16)
High trust	17.4% (n=26)
Average trust	45.6% (n=68)
Low trust	16.8% (n=25)
Very low trust	9.4% (n=14)

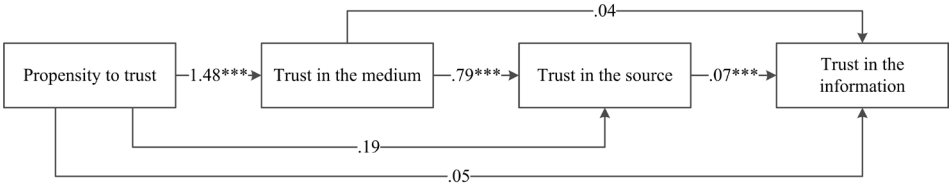

Figure 2: Cross section of the proposed layer model, showing unstandardized regression coefficients between all proposed constructs. Coefficients marked with three asterisks are significant at the 0.001 level; other coefficients were not significant.

Table 2: Motivations given for trust in the articles, coded in accordance with the 3S-model (Lucassen & Schraagen, 2011b).

Motivation	Percentage
Semantic features	11.5% (*n*=29)
Surface features	59.9% (*n*=151)
Source features	13.9% (*n*=35)
Other	14.7% (*n*=37)

and trust in the source as (sequential) mediators.

Basic regression analysis showed that the effect of propensity to trust on trust in the information was 0.20 ($p < .05$). However, when (either or both of) the two mediating variables were entered into the model, this direct effect became insignificant (0.05, $p = .56$).The total indirect effect was estimated at 0.08, with a 95% bias-corrected bootstrap (1000 samples) confidence interval of 0.03 – 0.16. Hence, trust in the medium and trust in the source mediated the effect of propensity to trust on trust in the information. Moreover, a model with only trust in the medium or trust in the source as mediating variable proved to be less valid, with a total indirect effect of respectively 0.05 (95% CI −0.02 – 0.16) and 0.01 (95% CI −0.05 – 0.08).

More light can be shed on the relationship between trust in Wikipedia and trust in information when considering the difference between high-quality and low-quality information. A median split on trust in Wikipedia was performed to distinguish participants with high and low trust in this source. Based on this split, we performed a repeated-measures ANOVA with article quality as a within-subject variable and trust in Wikipedia as a between-subject variable. A main effect of article quality and trust in Wikipedia on trust in the information was found, respectively $F(1, 147) = 7.45$, $p < .01$ and $F(1, 147) = 31.71$, $p < .001$. An interaction effect between article quality and trust in Wikipedia on trust in the information was only significant at the 10% significance level,

$F(1, 147) = 3.02$, $p = .08$. Visual inspection of the data suggested that users with low trust in Wikipedia were less influenced by article quality than users with high trust in Wikipedia. As expected, the same analysis applying a median split on propensity to trust and trust in the medium did not yield a significant interaction effect, $F(1, 147) = 0.37$, $p = .55$.

3.4 Motivations for Trust

A total of 131 of the 149 participants entered a motivation for their trust in the article on at least one occasion. This resulted in a total of 252 comments, which were categorized in accordance with the 3S-model (Lucassen & Schraagen, 2011b). Table 2 gives an overview of these comments.

As can be expected from a user group with limited domain knowledge on the topic at hand, most comments could be classified as referring to surface features (e.g. 'This article is well-cited'). However, semantic features (e.g. 'It appears to be historically correct, as far as my knowledge of the subject goes') and source features (e.g. 'Wikipedia has yet to fail me') were also mentioned as a motivation to trust the article. A remainder of 14.7% of the comments could not be classified in the 3S-model (e.g. 'I have no reason not to trust this particular article').

4. Discussion

In this paper, we propose a novel layer model of trust, with an inner core of trust in a particular piece of information, surrounded by trust in the source of the information, trust in the medium and propensity to trust in general. A mediation analysis on the results of the online experiment provided initial validation for this model. Moreover, a marginally significant interaction between trust in Wikipedia and trust in high-quality and low-quality information was found. The main contribution of this study is that the concepts of trust in a source, trust in a medium and propensity to trust in credibility evaluation are investigated in one, integrated study. This means that the presented model can be useful in explaining how these concepts are related to trust in information, and to each other.

Of course, the predictive power of each layer on the next is limited as numerous other aspects are likely to influence trust at the various levels as well (e.g. familiarity, information skills; Lucassen & Schraagen, 2011b). However, significant coefficients were found for each pair of layers. This means that we can draw the following conclusions:

- Trust in information is influenced by trust in its source.
- Trust in a source is influenced by trust in a medium.
- Trust in a medium is influenced by a propensity to trust.

Next to the correlation between trust in the source and trust in the information, a marginal interaction between trust in the source and trust in high-quality and low-quality information was found. In particular, this interaction can explain quite clearly how these two constructs are related. While it was only significant at the 10% level, it suggests an important difference between Internet users with a sceptical or trusting attitude towards Wikipedia. Earlier, we suggested that users with low trust in a source may skip it altogether, regardless of the information itself. We confirmed this behaviour in the experiment: participants with low trust in Wikipedia did not perceive any difference between high-quality and low-quality articles. This indicates a negative 'Halo effect' of the source on the information (Nisbett & Wilson, 1977); it is perceived differently (worse) because of characteristics of its source. Users demonstrating this behaviour are prone to undertrust, as they are likely not to use this source at all, even when the information quality is high.

On the other hand, participants with high trust in Wikipedia did perceive a difference between high-quality and low-quality information. This means that, even though they had a positive attitude towards the source, they still considered the quality of the information itself. Hence, no evidence for potential overtrust based on trust in the source was found in this experiment.

These findings are not in line with an earlier study (Sundar, Knobloch-Westerwick, & Hastall, 2007), in which it was shown that low source credibility led to the accumulation of cues from the information itself. Of course, the context of that study (various online news sources) was quite different from this one (Wikipedia). Thus, the effect found in our study may be specific for the case of Wikipedia. An alternative explanation is the extent to which the source was found not to be credible. It is possible that, when the source credibility is perceived to be limited, more cues in the information are sought, but when perceived source credibility is below a critical value, it is discarded entirely.

Moreover, it should be noted that, although a statistical difference was found between trust in high-quality and low-quality articles, this difference was quite small. Several explanations can be given for this finding. First, the motivation of the participants was likely to be limited, leading to a quick, heuristic evaluation of credibility at most (Metzger, 2007). Also, the categorization of the Wikipedia Editorial Team was taken as a measure for

quality. However, articles of lesser quality are not necessarily less credible, as the editorial team predominantly judges how far each article is from a distribution-quality article. In other words, completeness is a dominant factor for the editorial team, but this does not necessarily play a central role in the credibility evaluations of our participants.

Surprisingly, the link between trust in the source and trust in the information itself proved to be rather weak in this study. This finding seems to contradict much of the literature on the topic of source credibility, which mostly suggests that this is in fact a very strong relationship (Chaiken, 1980; Chaiken & Maheswaran, 1994; Sundar, Knobloch-Westerwick, & Hastall, 2007; Kelton, Fleischmann, & Wallace, 2008). Two explanations can be given for the lack of a strong correlation between trust in the source and trust in the information in this study.

First, the participants in this study were asked to evaluate multiple articles, which had one common characteristic, namely its source. This means that the participants were able to compare the articles with each other. In such comparisons, it is of no use to consider the source of the information, as this is a constant. The notion that the participants in this experiment indeed only made limited use of source cues is also supported by the percentage of comments (Table 2) regarding source features, which was rather low in this study (~14%). An earlier study (Lucassen & Schraagen, 2011b) featuring only one stimulus article yielded a much larger percentage of comments on the source (~30%). In contrast, in a think-aloud study with 10 stimulus articles (Lucassen & Schraagen, 2010), no utterances on the source of information were recorded at all. Hence, the ability to compare various articles could have diminished the influence of source cues.

Second, the particular source used in this study may have led to a limited influence on trust in the information. As already noted in the Introduction, it is problematic to consider Wikipedia as one single source in the traditional sense (Lucassen & Schraagen, 2011b). The information quality heavily varies between articles and over time, which makes the source credibility of Wikipedia a poor predictor for information credibility. Critical participants may have been aware of this, and thus attributed less value to source credibility. Some evidence of particularly critical participants can be found in the open-ended motivations, for instance in comments such as 'I don't really trust Wikipedia because someone from the public can make changes to the topic or article' and 'Wiki is an open source; anyone can comment on it and a slight change in the wording can cause misguidance.' This reasoning, as illustrated by these examples, may thus mean that at least some participants were aware

of the limited transfer of source credibility to information credibility in this particular context, which also may have led to a limited use of source cues.

No influence of trust in the medium or propensity to trust on trust in high-quality and low-quality information was found. This is in line with the hypothesized research model, as these constructs are more distant from trust in information.

A strong tie was found between trust in the Internet and trust in Wikipedia. This can partly be explained by the fact that, as opposed to the other layers, very similar questions were used to measure the two constructs (mostly only replacing 'the Internet' by 'Wikipedia'). However, we reckon that the relationship between these two is in fact among the most powerful, as Wikipedia is one of the most visited sites on the Internet (Alexa Top 500 Global Sites, n.d.). As stated (Dutton & Shepherd, 2006), trust in the Internet is largely built on experience with this medium. The prominent place Wikipedia has online means that its influence on trust in this medium is large. An interesting follow-up question is whether this relationship is equally strong in other contexts, such as television, newspapers or other printed materials.

Propensity to trust had a large influence on trust in the medium. This supports the notion that trust is specified from dispositional trust to more case-specific trust when needed (Merritt & Ilgen, 2008). However, since credibility evaluation in this context is largely heuristic, the disposition still has an influence on trust in information, albeit limited. It is expected that, in situations where the perceived need for credible information is higher (e.g. health information, financial transactions), propensity to trust is less influential, as trust is better calibrated to the actual credibility of the information as a result of a more profound evaluation of credibility (Metzger, 2007).

4.1 Limitations

Only one medium (the Internet) and one source (Wikipedia) were taken as a case study to demonstrate the validity of the research model in this experiment. Future research in this direction could utilize the same model, but with different media and/or sources to verify its validity in other contexts.

The interaction effect found between trust in the source and quality of the information on trust in the information was not significant at the customary 5% level. However, the trend found in this experiment suggests a larger risk of undertrust in Wikipedia than overtrust.

More research on the effect of the source of information on trust should confirm whether this is actually the case.

In this experiment, trust on various levels was measured through (partially validated) questionnaires. In future research, attempts should be made to manipulate trust more systematically, in order to rule out the effect of potentially confounding variables (e.g. age, Internet experience).

We cannot rule out the possibility that the order of the experiment (specifically the general questions on trust after the administration of the stimuli) had an influence on the answering of the questions. However, we are convinced that a reverse order in which the general questions would have been presented before the stimuli would have influenced the answering of the questions regarding the stimuli to an even larger extent, as this would have primed the participants on the issues of source, medium and propensity to trust. Future research could completely preclude this potential issue by counter-balancing the order of questions.

Finally, the results found here may not generalize to the entire population of Wikipedia (or Internet) users, as the sample is demographically biased towards European and North-American males. Gender differences in trust in Wikipedia (Lim & Kwon, 2010) and in general (Feingold, 1994) have been shown before. Moreover, six different articles were used as stimuli in this experiment.Although we attempted to rule out specific effects of this selection (e.g. very short or long articles), we cannot be sure that exactly the same effects are found when other articles are selected.

5. Conclusions

In this study, we proposed a novel model of trust. In this model, trust in information is influenced by trust in its source, which is in turn influenced by trust in the medium of this source. Moreover, trust in the medium is influenced by the user's propensity to trust. An online quasi-experiment has provided a first validation in the context of Wikipedia (as a source) and the Internet (as a medium). Moreover, some evidence for potential undertrust in Wikipedia was found, as participants with low trust in this source disregarded the presented information, without considering its actual quality in their credibility evaluation. No evidence for potential overtrust was found, as participants with

high trust in Wikipedia were still influenced by the quality of the presented information, rather than having blind faith in this source.

The proposed layer model can serve as a framework for future studies on the role of propensity to trust, trust in a medium and trust in a source in credibility evaluation, for example in other contexts than the Internet or Wikipedia.

Acknowledgments

The authors would like to thank Chris Kramer and Tabea Hensel for their valuable contributions to this study.

Appendix

Questionnaires used in the experiment

Propensity to Trust

- Regarding the intentions of others I am rather cynical and sceptical.
- I believe that you will be used by most people if you allow them to.
- I believe that most people inherently have good intentions.
- I believe that most people, with whom I have dealings, are honest and trustworthy.
- I become distrustful when someone does me a favour.
- My first reaction is to trust people.
- I tend to assume the best of others.
- I have a good deal of trust in human nature.

Trust in the Internet

- When you are looking for information, how often would you use the Internet as opposed to offline sources?
- What do you think is the credibility of the Internet?
- How much do you trust the institutes and people 'running the Internet'?
- How much confidence do you have in the people with whom you interact through the Internet?
- If you are in need of information, how confident are you that you can find it on the Internet?
- How well do you think your privacy is protected on the Internet?

Trust in Wikipedia

- When you are looking for information, how often would you use Wikipedia as opposed to other sources?
- What do you think is the credibility of Wikipedia?
- How much do you trust the institutes and people 'running Wikipedia'?
- How much confidence do you have in the people who add information to Wikipedia?
- If you are in need of information, how confident are you that you can find it on Wikipedia?
- How large do you think the risk of getting inaccurate information on Wikipedia is?

Chapter 7

Supporting Online Credibility Evaluation: A Wikipedia Case Study

In this study, we investigate the application of decision support in credibility evaluation of online information. Two important trade-offs are considered, namely the choice between user-based or automated support, and in case automated support is preferred, a complex support system (high performance, but hard to understand) or a simple support system (low performance, but easier to understand). Three simulated systems are evaluated through an experiment ($N = 72$) in the domain of Wikipedia. The results are largely in favor of the user-based and complex automated support. However, both have pitfalls that need to be considered. User-based support is very dependent on the number and credibility of the voters, which may be limited in dynamic environments. Complex automated support may suffer from inappropriate reliance, as users do not understand how it works. This study has applications in a broad field of online information services, especially for websites on which the credibility varies between pieces of information, or over time.

In preparation by Lucassen, T. & Schraagen, J. M. (n.d.).

1. Introduction

The rapid rise of the World Wide Web has resulted in a world in which anyone can access almost all information imaginable. However, in contrast to the pre-Internet era, the credibility of information is less certain. The problem has become even larger with the introduction of Web 2.0, in which anyone can easily make information available to the general public. These developments imply that Internet users should always consider the credibility of information they encounter. However, credibility evaluation has been proven to be a difficult task (Lucassen, Muilwijk, Noordzij, & Schraagen, 2013; Lucassen & Schraagen, 2011b), which is often neglected by Internet users (Walraven, Brand-Gruwel, & Boshuizen, 2009).

A potential solution for the problems that users have with the large variations in quality of online information, is to support them in their credibility evaluations. Numerous websites (such as *amazon.com* or *imdb.com*) already feature reputation systems, through which visitors can express their opinion on a topic of interest. Examples of such topics are other users, goods for sale, or pieces of information. Opinions on each topic are aggregated and presented to subsequent users, typically by a number of stars out of a total of five. Following this solution, the trust that a user has in information, is shifted to trust in other users, which in turn have indicated their trust in the topic (information) by means of a rating. This principle is founded on the philosophy of "wisdom of the crowds" (Surowiecki, 2005); the many are smarter than the few. However, this philosophy has been demonstrated to be fragile and potentially prone to bias (Lorenz, Rauhut, Schweitzer, & Helbing, 2011).

An alternative approach to improve credibility evaluations of Internet users is to develop an automated decision support system (DSS), as opposed to the user-based approach of a reputation system. In such an approach, the influence of the user is ruled out by the use of an automated algorithm which calculates a value for trust. This removes a potential source of bias due to unclear motives of voting users. However, credibility evaluation is a complex task involving many aspects, without a clear outcome in terms of a generally accepted truth. This implies that the level of complexity of algorithms used by a DSS is of importance; a trade-off between complexity and understandability of the support has to be made (Lee & See, 2004). More complex systems may yield better results in terms of accuracy and efficiency but less user acceptance because they may fail to understand how

the system works. In contrast, less complex systems are more easily understood by the users, resulting in better acceptance and more appropriate reliance but yielding poorer performance.

In this paper, we explore reputation systems and DSS of high and low complexity as support tools to enhance credibility evaluation online. The online encyclopedia Wikipedia is used as an example environment because of the large variations in information quality. The influence of three different simulated support tools (one reputation system and a simple vs. complex DSS) on Wikipedia users is explored in a lab experiment.

The rest of this paper is structured as follows. We first start with a literature review on online credibility evaluation, followed by a discussion on the application of reputation systems or automated DSS in credibility evaluation. After this, we discuss our methodology to evaluate the three simulated support tools, followed by the results of the experiment and a discussion about its implications for the design of a support tool for credibility evaluation.

1.1 Credibility Evaluation

Credibility and trust are often confused in the literature. In this study, we refer to credibility as a property of information and to trust as a property of the reader (user) of the information. Credibility of information can be described as believability (Fogg & Tseng, 1999). In cognitive psychology, credibility is split into two key elements, namely trustworthiness and expertise of the author of information. In order for information to be credible, its author must be well-intentioned (trustworthy) and knowledgeable on the topic at hand (expertise).

Based on the credibility of information, a reader may decide whether to trust or not trust it. This decision always involves a certain risk, as one can never be entirely certain of the credibility of information (Kelton, Fleischmann, & Wallace, 2008). Trust involves the willingness to take this risk and depends on the credibility of information. This is for instance expressed by the choice to use the information (e.g., by taking actions based on it). In order to reduce the risk taken by trusting information, credibility evaluations may be performed by users. In such evaluations, users search for cues in the information that may be indicators of credibility (Fogg, 2003). Examples of such cues are the length of the text, the number of references, or the presence of pictures (Lucassen et al., 2013; Lucassen & Schraagen, 2011b).

Trust in information is especially critical in online environments. In contrast to the pre-internet age, anyone can easily make information available to everyone, resulting in a rising amount of poor-quality information. Professional gatekeepers (e.g., editors, publishers) are mostly not available anymore (Flanagin & Metzger, 2007) which means that end-users have to evaluate credibility themselves. This may be problematic as they often lack training on this task.

How people cope with the task of evaluating credibility online has been a topic of research for some time now. Fogg (2003) proposed that each element of a website has a certain prominence to each user. When an element (e.g., author information) is prominent, a user may have a certain interpretation of what this means for the credibility of the web page. The interpretations for all prominent elements are summed up to form a credibility judgment. Prominence and interpretation are likely to vary substantially between users, which means that the same piece of information may have a very different perceived credibility for different users.

This notion was extended by Lucassen and Schraagen (2011b), who proposed that different properties of the user may lead to different features of the information being incorporated in credibility evaluations. Information may be considered on two levels, namely the semantic level (e.g., factual accuracy, neutrality) or the surface level (e.g., writing style, references). The semantic level can only be considered when the user has a level of domain expertise on the topic at hand. When domain expertise is not available, the user may still evaluate credibility in a meaningful way, namely by considering surface properties of the information. However, this requires a level of information skills of the user. Next to these two strategies, a user may also consider the source of information (e.g., the website the information was found on), thereby relying on his experience with this particular source.

It has been shown that by following these strategies, certain user groups are able to distinguish high quality from low quality information. For instance college students and PhD candidates are able to distinguish good Wikipedia articles from poor Wikipedia articles, regardless of their prior knowledge of the topic, whereas high school students are not able to do so (Lucassen et al., 2013). Not all Internet users are thus expected to have high information skills. In (Lucassen & Schraagen, 2011b), it was shown that novices in the area of automotive engineering could not detect factual errors in Wikipedia articles on car engines. Hence, their trust was not affected by these errors. Experts were affected by the errors, albeit to a limited extent.

An explanation for the differences in credibility evaluation between users can be found in the dual-processing model of web site credibility evaluation (Metzger, 2007). It was proposed that the extent to which credibility is evaluated depends on the motivation and ability of the user. When a user is not motivated at all, no evaluation will be performed. When the user has a motivation (e.g., a perceived risk of poor information), the choice between a heuristic and systematic evaluation depends on the ability of the user to evaluate. Only when the user has high capabilities, a systematic evaluation is performed.

Thus, even the select group of Internet users who possess the appropriate information skills to evaluate the credibility of information still may not perform such evaluations, as they may lack the motivation to evaluate. This was also shown by Walraven, Brand-Gruwel, and Boshuizen (2009); in an interview, high school students were able to name several aspects they would consider when evaluating credibility, thereby showing some evaluation capabilities. However, when they were actually searching for information, none of these aspects were taken into account. The discrepancy between the aspects named and the actual evaluation behavior can partly be explained by a lack of motivation. An alternative explanation is working memory overload (cognitive load theory; Chandler & Sweller, 1991), which would also be an indication for a lack of ability to evaluate.

1.2 Reputation Systems

With the current trend of user-generated content on many websites (e.g., Wikipedia), it would seem obvious to also take a community-driven approach in credibility evaluation support. The traditional method for such an approach is a reputation system.

An extensive literature review on online reputation systems is given by Jøsang, Ismail, and Boyd (2007). They argue that with such systems, trust based on direct interaction with the object at hand is replaced by derived trust from a referral (e,g., opinions of other users). However, users may still consider their own judgment of the credibility of information, in addition to the advice from others. This principle, compared to trust in information without a reputation system, is shown in Figure 1.

Resnick, Zeckhauser, Friedman, and Kuwabara (2000) identified three requirements for a reputation system to perform effectively. First, the entities (or objects of interest) which gain reputation should have a long life-expectancy. It takes time to build a reliable reputation as this is likely to rise with the number of opinions. If an entity seizes to exist or changes before an appropriate number of opinions is gathered, a reputation system will

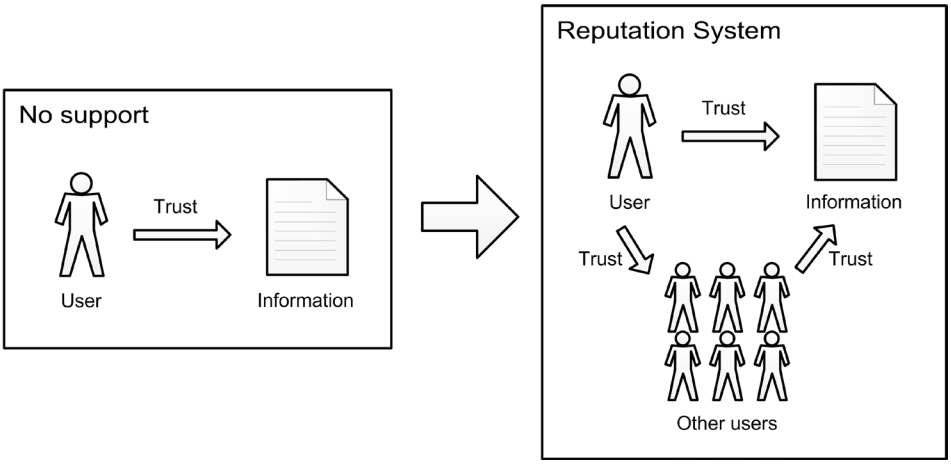

Figure 1: Trust in information, with and without a reputation system. When a reputation system is present, trust in the information is shifted to trust in others.

have very limited added value. Second, a reputation system should capture the opinions about current interactions to show in future interactions. Third, the feedback of the reputation system should be used to guide trust decisions. In other words, if trust is not influenced by the system, it has no use.

A proposal to implement a reputation system on Wikipedia has been made by Korsgaard and Jensen (2009). They suggested some small changes in the MediaWiki software (on which Wikipedia runs) to facilitate user ratings. These ratings are attached to revisions of an article, because credibility is likely to improve (or at least change) as an article evolves as the result of edits to its contents. User ratings on previous revisions of an article have to be discarded in the current revision, because weaknesses in the article which led to lower ratings in a previous revision might have been dealt with in the current revision. Currently, Wikipedia is developing and testing such a reputation system . The "Article Feedback Tool" (Wikipedia:Article Feedback Tool, n.d.) can be used to rate trustworthiness, objectivity, completeness, and writing style, and is visible at the bottom of most articles.

The deletion of ratings of previous revisions is a potential drawback of reputation systems in dynamically evolving systems such as Wikipedia. As information on such websites is updated regularly, the period in which information remains unchanged is limited. This is a direct violation of the first requirement of Resnick et al. (2000); only a small number of users will be able to submit a rating before information is updated. The chance of

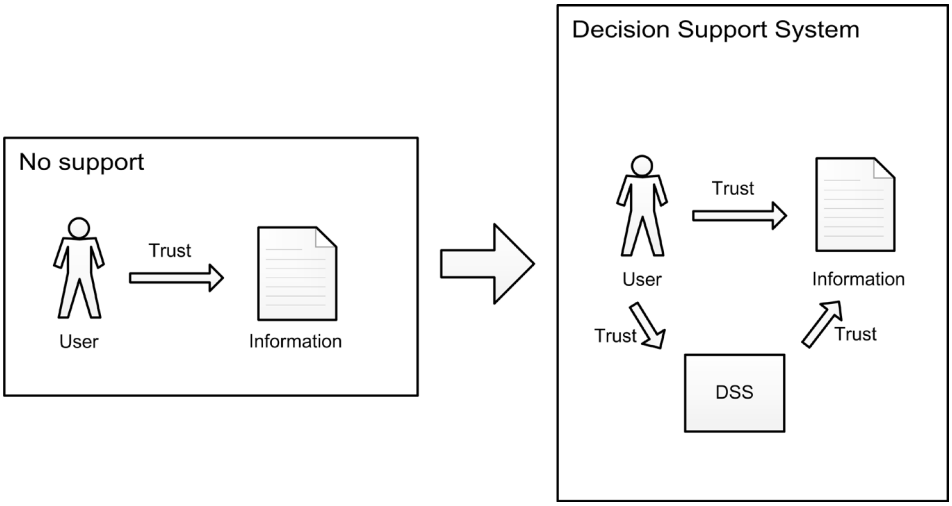

Figure 2: Trust in information, with and without a decision support system. When a reputation system is present, trust in the information is shifted to trust in the support.

encountering information with a rating based on a low number of visitors or no rating at all may thus be quite high. This limits the reliability and usefulness of such systems, as we suspect that reliability heavily correlates with the number of ratings of previous visitors. Nevertheless, this method of earlier visitors voting on the credibility of articles corresponds very well with the collaborative nature of Web 2.0. By implementing such a system, users are not only involved in contributing information, but also in rating its credibility.

A second threat for the efficiency of reputation systems as a support tool for credibility evaluation lies in the "wisdom of the crowds" principle. Galton has already shown in 1907 that the knowledge of many novices may outperform that of a few experts (Galton, 1907), and recent success cases have been listed in (Surowiecki, 2005). However, this principle has also been shown to be very fragile (Lorenz et al., 2011). In the original experiments on this principle, participants were not aware of the answers of other participants. Lorenz et al. (2011) demonstrated that when they are shown the other answers, they tend to adapt their own answer to the current average. This has significant implications for online reputation systems. If users are able to see the average voting before they vote themselves, they are very likely to adapt their own opinion to the same range, thereby altering their original opinion. This may lead to overly optimistic ratings when the first few opinions were positive, or to overly pessimistic ratings when the first few were negative.

1.3 Decision Support Systems

In order to overcome the potentially negative influences of the user aspect in reputation systems (low number of votes, biases in voting), a different approach in supporting users in credibility evaluation may be taken, namely automated support. Instead of considering user opinions, algorithms could be used to determine the credibility of information, which can in turn be communicated to the user, as shown in Figure 2.

A vast body of research is available on (automated) DSS (see Lee & See, 2004, for a review). One returning topic is the reliance of the user on the support. When a support system is not 100% reliable, which is mostly the case in real-life systems, the user should rely appropriately on the advice of the system. Appropriate reliance implies that the level of trust in the support system matches its actual performance. Two pitfalls can be identified; first, a user may overly rely on the support. In this case, trust in the support exceeds the actual capability of the support, leading to misuse (Parasuraman & Riley, 1997). Second, a user may under-rely on the support. This happens when the capability of the system exceeds trust of the user, leading to disuse.

Lee and See (2004) have proposed a model which explains how a reliance action is formed. First, a user assimilates the information needed to form a decision. Second, the influence of this information on trust is evaluated. Third, a reliance intention is formed and fourth and last, a reliance action is taken.

An important challenge lies in the information assimilation stage of this process. The information may be gathered from the context in which the support is used (being individual, organizational, cultural or environmental) and from the support system itself. In order to optimize appropriate reliance, Lee and See (2004) advise that support systems should be made easy to understand, using simple algorithms clear to the user. In this case, in situations where the support fails, the user also has the opportunity to understand why, leading to more appropriate reliance. The advantages of understandability has also been demonstrated by Ye and Johnson (1995). Participants in their experiment had a greater belief in the truthfulness and reasonableness of the support system after they were given a clear explanation of the advice.

On the other hand, more complex support systems may perform better, especially in complex task settings such as credibility evaluation, where many factors may contribute to credibility in various ways. However, such support systems may be harder to understand

by the user, if at all. Lee and See (2004) name this challenge the trade-off between *trustable* and *trustworthy* advice. The first supplies understandable information about how it functions, which makes the support trustable, whereas the latter performs better, and is thus more trustworthy.

When considering DSS in the domain of online credibility evaluation, an important difference with traditional DSS can be seen. In many cases, support systems are designed to assist professionals, who are motivated to perform a very specific task to the best of their abilities. Examples are support for air traffic control (Hilburn, Jorna, Byrne, & Parasuraman, 1997) or in the naval domain (Van Maanen, Lucassen, & Van Dongen, 2011). However, in online credibility evaluation, not much is known about the user. Motivation levels are questionable (Metzger, 2007), as are information skills (Lucassen et al., 2013, Lucassen & Schraagen, 2011b). This has the following important implications for the design of a support system:

1) Because of the widespread user group (Internet users), extensive training with a support system is virtually impossible. Therefore, the support should be very easy to use.
2) Motivation may be (very) limited, using the support should not take too much time and effort. Ideally, users should be even faster in evaluating credibility with a DSS than without.
3) The issue of credibility evaluation may not be salient to all Internet users. The mere presence of a support system may influence this attitude of the user; why would there be a support system if the information is always credible?

A good example of a highly relevant domain for the application of credibility evaluation support systems is Wikipedia. The overall information quality of this online encyclopedia is very high (Giles, 2005). However, its information quality fluctuates heavily between the articles (Denning, Horning, Parnas, & Weinstein, 2005), due to its open-editing model. A few attempts have already been made to implement automated support systems in the domain of Wikipedia. An example of a support system focusing on understandability is *WikipediaViz* (Chevalier, Huot, & Fekete, 2010).This is a system for Wikipedia that presents various statistics about an article (e.g., number of words, contributors and internal links) to the user. Because of the fact that only raw statistics are presented, without further processing by the support, it is very clear to the user what this system actually does. However, for at least some of these statistics, it may be unclear what their

relationship to credibility is. For instance, as far as the number of authors is concerned, it is unknown whether articles with a lot of authors are to be trusted more or less than articles with fewer authors.

Adler et al. (2008) have introduced a more sophisticated (and complex) algorithm to support credibility evaluation called *WikiTrust*. Based on ideas proposed earlier (Cross, 2006; McGuinness et al., 2006), this support system calculates a trust value on a single word basis. First of all, the credibility of a particular word is based on how long this word has been in the article. When a new word is added, it is assigned the credibility of its author, which in its turn is based on the average credibility (survival duration) of his additions. The credibility of each word rises over time as the article is edited but this word is kept unchanged. Credibility is visualized by coloring the background of each word (instead of the words themselves) in a shade of orange. Dark orange indicates low credibility; white implies high credibility (the actual algorithm is slightly more complex than described here).

An indication of how well such a complex support system could perform in terms of user acceptance can be found in Lucassen and Schraagen (2011a). They showed that college students have difficulties incorporating the extra information by *WikiTrust* into their own credibility evaluations. It can be argued that this is due to differences between the mental model of trust of the participating students and the features used by the support. This notion is supported by Hilligoss and Rieh (2008), who claim that the user's personal definition of trust (e.g., features involved) is highly influential on the way credibility is evaluated by this user. When a support system is based on features that are not deemed important for credibility by the user, it is likely not to be accepted. Hence, next to understandability of support, agreement with the user about the features used by the support is a vital requirement for user acceptance.

1.4 Research Questions

The support systems for credibility evaluation (in this case in the domain of Wikipedia) discussed in the previous section seem to be primarily technology-driven. They focus on features that can be measured automatically and bear some relevance to the credibility of information, at least in the eyes of their creators. Following this approach, the way in which the user interacts with these systems is largely neglected. In this paper, we consider various design options from the perspective of the user. Following the considerations

presented in the preceding sections, we formulate the following research questions:

1) Should the development of support systems for online credibility evaluation focus on user-based or automated systems?
2) Should the development of automated support systems for online credibility evaluation aim at understandability or performance?

We attempt to answer these questions by performing an experiment in which we ask users to evaluate the credibility of Wikipedia articles while confronted with one of three different support systems. The first is a user-based reputation system, whereas the latter two are respectively a simple and a complex automated support system.

2. Method

2.1 Participants

A total of 72 college students participated in an experiment (20 male, 52 female). Their mean age was 20.3 (SD = 4.1). 48 students were Dutch, 24 were German. All participants were familiar with reading texts in the English language used in this experiment as they were following an academic education, given partly in this language. All participants reported to be acquainted with Wikipedia, on average having 6.3 (SD = 1.7) years of experience with the online encyclopedia. Most participants were regular Wikipedia users; frequency of use was reported as every day (4%), every week (58%), every month (32%), or every year (4%). Only one participant indicated to never use Wikipedia. All participants were able to explain how Wikipedia works in their own words. Seven participants had some experience in editing Wikipedia articles.

2.2 Task

The participants performed the Wikipedia Screening Task (Lucassen & Schraagen, 2010). In this task, the participant is asked to evaluate the credibility of a presented Wikipedia article. The strategy to perform this task is not specified; participants were allowed and encouraged to invent their own methods. After evaluating an article, the participant was asked to rate the credibility on a 7-point Likert scale with only the end-points labeled. A motivation to their judgment was also asked, along with their familiarity with the topic of the article and their confidence in their judgment.

2.3 Materials

2.3.1 Articles Used

Eight articles were used in this experiment, all obtained from the English Wikipedia. Offline versions were created, in which any direct cues of credibility (such as warnings about neutrality or citations) were removed. This was done to ensure that such cues had no influence on the judgments of the participants.

In order to ensure that the articles used in the experiment were a representative sample of Wikipedia, they were selected from various areas of interest of the encyclopedia. This also eliminates potential effects of particular topics on trust in the information or in the support systems. Topics on untrustworthy persons for instance, may have a negative impact on perceived credibility, even when they have been written in a very concise and neutral manner. Wikipedia features eight main categories on its main page; one article was selected from each category. These can be found in Table 1.

A second important consideration was article quality. The Wikipedia Editorial Team (Wikipedia: Version 1.0 editorial team, n.d.) has assessed most articles and categorized them into one of seven categories, ranging from stub to featured article. In this study, the articles should be substantial in the sense that they provide enough material (cues) for the participants to base their decision on. Examples of these cues are the presence of references, pictures and a reasonable amount of text (Lucassen et al., 2013; Lucassen & Schraagen, 2010). On the other hand, in order for support systems to be helpful,

Table 1: Articles used in the experiment.

Article	Topic	Category	# words	# references	# pictures
1	Cha-cha-cha (Dance)	Arts	1127	9	3
2	Lied Glacier	Geography	1139	10	1
3	Isabelle Adjani	Biography	983	12	1
4	Watt's Curve	Mathematics	651	7	2
5	Medieval Warm Period	History	1546	39	3
6	Critical Mass	Society	1848	41	4
7	Alpha Decay	Science	1265	7	3
8	Linear Motor	Technology	1853	8	5

ambiguity in credibility should exist. In other words, there should be potential reasons for the participants not to trust the information in the articles. One category as defined by the WPET seems particularly suitable to be used in this experiment, namely the start-class. A short description of an article in this category is: "An article that is developing, but which is quite incomplete and, most notably, lacks adequate reliable sources." Therefore, all articles in this experiment were obtained from the start-class category.

Considerations have also been given to other issues such as controversy and familiarity. The participants were Dutch and German psychology students, so topics such as psychology or Dutch or German history were not used. Instead, all articles were assumed to be on relatively unfamiliar topics, replicating normal expected Wikipedia usage (searching for new information).

2.3.2 Support Systems

Three support systems were simulated in the experiment. The support systems were not actually functional, but the stimuli and explanations presented to the participants gave the impression that they were. In order to rule out any influence of the manner in which the advice was presented to the user, this was kept constant for all systems. A design assumed familiar to most participants was selected, namely a rating expressed by a number of highlighted stars with a maximum of 5, accompanied with a more detailed numerical representation with one decimal underneath the stars. This presentation is similar to popular websites such as imdb.com and amazon.com. Examples for each support system are given in Figure 3. This rating was presented in the right upper corner of the Wikipedia article.

The first system that was investigated represented a user-based reputation system, called *Trustworthiness User Rating*. The presented rating was based on votes of prior visitors. The number of votes was kept constant around 90. Other websites regularly feature much higher numbers of votes, but it is unrealistic to expect such a high number of votes for Wikipedia, because of the dynamical environment of this website. The number of votes was shown next to the numerical rating underneath the star rating.

Figure 3: Presentation of all three support systems.

Secondly, the *Decision Rules Rating* support system was introduced to act as a simple, understandable automated support system. The presented rating was based on three simple decision rules, namely the number of words, the number of references and the number of pictures. While evaluating these three features from the article is clearly not the same as evaluating credibility, these features were mentioned regularly by participants evaluating credibility of Wikipedia articles in earlier research (Lucassen et al., 2013; Lucassen & Schraagen, 2010). The predictive powers of these characteristics have also been shown in other studies. Word count has been shown to be a very successful predictor of information quality (Blumenstock, 2008). The number of images and characters (instead of words) significantly differ between featured and random articles (Stvilia, Twidale, Gasser, & Smith, 2005). Dondio, Barett, Weber, and Seigneur (2006) also incorporated the number of references in their prediction model of information quality. Thus, next to the assumption that the proposed characteristics are meaningful to the participants, they also play important roles in prediction models. The counts of the three features were shown next to the numerical rating underneath the star rating to provide a rationale for the given advice.

Table 2: Instructions for all three support systems. Note that these instructions are translations, as the experiment was performed in Dutch.

Support System	Instructions
A Trustworthiness Rating	This system is similar to well-known recommender systems, for example as used by amazon.com. Previous visitors to this page have submitted a credibility judgment; the average is represented by the number of stars at the top right corner of this article.
B Decision Rules Rating	This system uses a few simple decision rules: - Longer article means more trustworthy - More references means more trustworthy - More images means more trustworthy All rules are equally important. The result of combining them is represented by the number of stars at the top right corner of this article.
C Adaptive Neural Network Rating	This system uses an adaptive neural network that calculates credibility on the basis of numerous characteristics (such as characteristics of the authors, multiple details of the edit history). The result is represented by the number of stars at the top right corner of this article.

Finally, the *Adaptive Neural Network Rating* support system represented a complex automated support system. The exact workings of this system were not specified (actually, even non-existent), but the explanation stated that the advice was based on numerous features of the article, such as the edit history and the authors. The name was invented such, that it suggested a complex, sophisticated algorithm generating the advice. A similarly complex (real) system is *WikiTrust* (Adler et al., 2008); most Wikipedia users will most likely not be able to fully comprehend the algorithm behind this support system. Since we intended the user not to be entirely informed about how the system worked, no further details to the rating were given.

The instructions for each support system that the participants in our experiment received are listed in Table 2. From here on the Trustworthiness User Rating is referred to as Support A, the Decision Rules Rating as Support B, and the Adaptive Neural Network Rating as Support C.

Seven versions of each article were generated: one without support, and two for each of the three support systems, namely one with positive advice and one with negative advice. Positive and negative advice were given both to investigate whether the various support systems could influence the ratings of the participants in both directions. Positive advice was reflected by a rating of around 4.5 out of 5 stars; negative advice was reflected by about 1.5 out of 5 stars as it is not customary for star ratings to be able to score less than one star.

2.4 Procedure

After giving informed consent, instructions were given to the participant verbally and on the computer screen. The entire experiment was performed digitally using LimeSurvey software (Schmitz, 2012). The participants were told that they were free to ask questions at any time. First, a few questions on the demographics of the participants were asked, along with questions on their experience with Wikipedia. After these questions, the first Wikipedia article was presented. The order of the eight articles was the same for each participant (as given in Table 2), whereas the order of the support was balanced between subjects using a Latin square, based on the order N (no support), A (Trustworthiness User Rating), B (Decision Rules Rating), and C (Adaptive Neural Network Rating). Two articles were subsequently presented in each condition, one with positive advice and one with negative advice. The order of positive and negative advice was randomized. Instructions on each support system were given prior to the first article featuring that

support. Additionally, an instruction sheet with all support systems was supplied so the participants could re-read the instructions whenever they needed. After assessing all articles a few control questions on the manipulations and opinions on the support systems were asked. The duration of the experiment was about 60 minutes.

2.5 Dependent Variables

The following dependent variables were recorded in the experiment.

Trust in the information. After each article, the participants were asked to indicate the credibility of the article on a 7-point Likert scale.

Confidence in judgment. After each article, the participants were asked to indicate how confident they were about their trust judgments on a 7-point Likert scale.

Familiarity with the topic. After each article, the participants were asked to indicate how much they already knew about the topic of the article beforehand on a 7-point Likert scale.

Trust in support. After each article featuring one of the support systems, the participants were asked to indicate the perceived credibility of the provided support system on a 7-point Likert scale, along with a motivation to their answer through an open-ended question.

Opinion on the support. After viewing all articles, the participants were asked to compare the support systems with each other and indicate their opinion on several aspects on 7-point Likert scales. These aspects were:

- Usefulness (Is the support system useful?)
- Time needed (Does the support system help to form a judgment more quickly?)
- Quality (Does the support system help to form a better judgment?)
- Effort (Does the support system help to reduce the effort it takes to form a judgment?)
- Basis of judgment (To what degree is your judgment based on the advice of the support system?)

Most of the answers of the participants were given through (7-point) Likert scales with only the end-points labeled. These scales are assumed to measure at an ordinal level rather than an interval level (Jamieson, 2004). Therefore, non-parametric Wilcoxon Signed Rank Tests were used in the analysis of the data.

3. Results

3.1 Trust in the Information

The level of trust in each individual article presented in the experiment when no support was given is shown in Table 3.

As stated in the method section, all articles were selected from the same quality class. Naturally, this does not mean that the credibility is exactly the same for each article. The results show that average trust in each article ranged from 4.3 to 5.6 with an total average of 5.1.

Table 4 shows the credibility ratings for all support conditions in comparison to the no support condition. Support system A was only influential with negative advice (positive: $Z = 1.13$, $p = .257$; negative: $Z = 4.15$, $p < .001$) whereas support system B only influenced the trust judgments when giving positive advice (positive: $Z = 2.61$, $p < .01$; negative: $Z = .69$, $p = .489$). For support system C, both positive and negative advice influenced the trust judgments (positive: $Z = 3.31$, $p < .01$; negative: $Z = 2.50$, $p < .05$).

Table 3: Trust in each individual article without support. Average scores on 7-point Likert scales ranging from 1 (very untrustworthy) to 7 (very trustworthy) for all conditions. Standard deviations are given in parentheses.

Article	Topic	Trust
1	Cha-cha-cha (Dance)	5.6 (1.1)
2	Lied Glacier	5.6 (1.2)
3	Isabelle Adjani	4.3 (1.3)
4	Watt's Curve	4.7 (1.4)
5	Medieval Warm Period	5.4 (1.1)
6	Critical Mass	5.4 (1.1)
7	Alpha Decay	4.9 (1.6)
8	Linear Motor	5.2 (0.9)
	Average	5.1 (1.1)

Table 4: Trust in the information. Average scores on 7-point Likert scales ranging from 1 (very untrustworthy) to 7 (very trustworthy) for all conditions. Standard deviations are given between parentheses; scores which are significantly higher or lower than in the no support condition are marked with an asterisk.

	No Support	Support A	Support B	Support C
Positive advice	5.1 (1.1)	5.3 (1.4)	5.6 (1.2)*	5.8 (1.1)*
Negative advice		4.2 (1.4)*	5.3 (1.2)	4.6 (1.4)*

Table 5: Trust in the support. Average scores on 7-point Likert scales ranging from 1 (very untrustworthy) to 7 (very trustworthy) for all support conditions. Standard deviations are given in parentheses.

	Support A	Support B	Support C	All
Positive advice	4.6 (1.6)	3.8 (1.8)	4.8 (1.5)	4.4 (1.1)
Negative advice	3.8 (1.6)	2.6 (1.4)	3.8 (1.4)	3.4 (0.9)
Total	4.2 (1.6)	3.2 (1.7)	4.3 (1.6)	3.9 (1.1)

Table 6: Confidence in judgments. Average scores on 7-point Likert scales ranging from 1 (very high confidence) to 7 (very low confidence) for all conditions. Standard deviations are given in parentheses; scores which are significantly higher or lower than in the no support condition are marked with an asterisk.

	No Support	Support A	Support B	Support C
Positive advice	n/a	4.3 (1.4)	4.7 (1.3)*	4.7 (1.4)*
Negative advice		4.0 (1.5)	4.7 (1.3)*	4.2 (1.5)
Total	4.1 (1.3)	4.2 (1.4)	4.7 (1.3)*	4.4 (1.5)

The two support systems who were successful in influencing trust positively (B and C) were also compared to each other to assess the extent to which they raised trust. The same was done for the two support systems who influences trust negatively (A and C). No difference was found in trust in information with positive advice from support B or C, $Z = 0.84$, $p = .399$. Negative advice from support A had more influence on trust in the information than support C, $Z = 2.12$, $p < .05$.

3.2 Trust in the Support

Table 5 shows trust in the support systems. Positive advice from any of the support systems was trusted more than negative advice ($Z = 5.88$, $p < .001$). Support systems A and C were both trusted more than support system B (respectively $Z = 4.09$, $p < .001$ and $Z = 4.61$, $p < .001$), with no significant difference between them ($Z = .55$, $p = .583$).

3.3 Confidence in Judgments

Table 6 shows the confidence of the participants in their own judgments. Support system A did not significantly influence the confidence of the participants ($Z = .12$, $p = .902$). The participants' confidence in their own trust judgments was higher when support system B was present ($Z = 3.55$, $p < .001$). Support system C did not significantly influence the participants' confidence ($Z = 1.53$, $p = .126$). A more detailed analysis showed that positive advice from support system C did raise the confidence of the participants ($Z = 2.64$, $p < .01$).

Table 7: Opinions on the support. Average scores on 7-point Likert scales ranging from 1 (very untrustworthy) to 7 (very trustworthy) for all conditions. Standard deviations are given in parentheses.

	Support A	Support B	Support C
Usefulness	4.1 (1.8)	2.9 (1.6)	4.0 (1.8)
Speed of judgment	4.2 (1.8)	3.1 (1.8)	3.8 (1.9)
Quality of judgment	3.8 (1.7)	3.3 (1.8)	3.8 (1.7)
Effort of judgment	3.5 (1.8)	3.1 (1.8)	3.4 (1.7)
Basis of judgment	3.3 (1.5)	2.7 (1.6)	3.4 (1.7)

3.4 Opinions on Support

Table 7 shows the opinions of the participants on several aspects of the support. Usefulness, speed of judgment, quality of judgment, and basis of judgment were rated higher for support system A and C than B (all p's < .05) with no difference between them (all p's > .20). No effects were found for effort (all p's > .05).

3.5 Motivations for Trust in Support

Categorization of the open-ended questions on the motivation for the participants' trust in the support was performed to gain insight in the attitude towards each support system. Based on inspection of the motivations of the first few participants, several ad-hoc categories were created. Additional categories were added when needed. The categories, along with the number of motivations that were grouped into them, are shown in Table 8.

The motivations of 16 of the 72 participants (22%) were also categorized by a second rater. Based on the overlap, Cohen's kappa was calculated to ensure inter-rater reliability. The resulting value of .704 indicates a substantial agreement.

A motivation given for each of the three support systems was that the support did or did not match with their own judgment. The first was the main reason to trust the support, whereas the latter was a definite reason not to trust the support. Mismatches mostly occurred when negative advice was given.

Moreover, motivations which could not be interpreted by the experimenters were categorized as "Other/No motivation". Some participants filled in a motivation to trust the support which was actually a motivation to trust the article itself. Others were unclear in their wording, or provided no motivation at all.

Table 8: Motivations for trust in the support.

Motivation	Positive advice	Negative advice
Support A		
Assumptions about the credibility of the voters	28 (39%)	27 (38%)
Number of votes	15 (21%)	13 (18%)
Match between own judgment and advice	6 (8%)	7 (10%)
Mismatch between own judgment and advice	4 (6%)	10 (14%)
No rationales for the votes are given	2 (3%)	2 (3%)
Only the average opinion is given	2 (3%)	2 (3%)
Other/No motivation	15 (21%)	11 (15%)
Support B		
The features used are good indicators of credibility	5 (7%)	0 (0%)
The features used are poor indicators of credibility	38 (53%)	50 (70%)
Match between own judgment and advice	6 (8%)	3 (4%)
Mismatch between own judgment and advice	0 (0%)	0 (0%)
Match between advice and own observation of features	12 (17%)	11 (15%)
Other/No motivation	11 (15%)	7 (10%)
Support C		
The features used are good indicators of credibility	23 (32%)	15 (21%)
The features used are poor indicators of credibility	17 (24%)	12 (17%)
Match between own judgment and advice	8 (11%)	2 (3%)
Mismatch between own judgment and advice	1 (1%)	16 (22%)
It is unclear what the support actually does	5 (7%)	8 (11%)
Other/No motivation	18 (25%)	19 (26%)

Support system A was often not trusted because the participants were less sure about the credibility of the voters than of the information itself, rendering the support system less credible. This is remarkable, as our participants did not know who wrote the article on Wikipedia and who voted on it. Both the writers and voters are likely to stem from the same pool of users, namely the Wikipedia community.

A second motivation that was regularly observed was that the participants believed the number of votes (about 90) to be too low for reliable advice. After the experiment, participants were asked to indicate the minimum number of votes they thought would be needed for reliable advice. About half of their answers exceeded the available number of votes in our experiment.

Other motivations concerned the manner of presentation of the support (average only, no rationale for low/high votes).

Support system B was largely distrusted because the participants did not find the features used (length of the text, number of references and number of images) suitable indicators of credibility, although some participants thought they were good indicators. However, it was most often stated that they had nothing to do with the credibility of an article.

Furthermore, the participants regularly checked whether the number of references and pictures in the article actually matched the number displayed by the support.

Support system C yielded some remarkable motivations. The instructions for this support system were kept intentionally vague, but most of the participants filled in the blanks themselves. In other words, they came up with their own interpretation of how the support works. This resulted in most of the participants finding the support system very credible. However, there were still a considerable number of participants who found "characteristics of the author and details of the edit history" unsuitable indicators for credibility.

Moreover, some participants did not come up with a personal interpretation of how the support system worked. Instead, these participants indicated that it was unclear to them how the system actually worked, and therefore had little trust in it.

4. Discussion

In this study, we explored the application of decision support in credibility evaluation of online information. We focused on answering two questions in particular. First, we wanted to investigate whether it is more beneficial to introduce a user-based reputation system or an automated support system that does not require user involvement. Second, when one chooses to develop an automated support system, should the focus be on understandability or performance?

4.1 Simple vs. Complex Support

Before comparing automated support to the user-based reputation system, let us discuss the difference between both automated systems. The experiment performed in this study has provided several indicators to guide us in choosing between the various possibilities for support systems. We can consider the influence of the support on trust in the information,

trust in the support itself, the confidence of the participants in their judgment, and various other opinions on the support.

The results of the experiments seem to be mostly in favor of the complex support system. It was shown that the simple support system could only increase trust by giving positive advice, whereas the complex support could also diminish trust by giving negative advice. The latter is perhaps the most important influence of support, as this prevents overtrust in the information. Also, trust in the complex support was higher than in the simple support. Furthermore, usefulness, speed of judgment, and quality of judgment were higher for the complex support and thus the participants based their judgment more on that system.

The motivations of the participants could shed more light on why the complex support system performed better than the simple support system. Two motivations were recorded frequently, namely 1) the features used by the complex support were good indicators for credibility, and 2), the features used by the simple support were poor indicators. We first discuss the opinion on the features used by the complex support, followed by a discussion on the features used by simple support.

The observation that the complex support used very appropriate indicators of credibility seems counterintuitive at first glance. After all, the information about how the system works implied a complex algorithm, but was quite meager about the exact inner workings. So how could our participants find the indicators appropriate if they did not exactly know what these were? This phenomenon can be explained by a 'positivity bias' towards the support. It has been found that people have a general tendency to expect positive things from unknown people or objects when only little information is available (e.g., Mezulis, Abramson, Hyde, & Hankin, 2004). This has also been demonstrated in the domain of automated decision support by Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003), who found that students believed that the support would perform well when they were provided only very limited information about its reliability. This observation implies that the participants must have made their own assumptions about the inner workings of the support. Because of the meager information about the workings of the support system, it is likely that these assumptions stem from their own mental model of credibility (Cohen, Parasuraman, & Freeman, 1998; Hilligoss & Rieh, 2008) rather than the instructions. This could also be seen in the motivations of the participants, which often contained interpretations not found in the instructions or in the support system itself. Having thought up these interpretations themselves, this means that it is likely

that they largely agreed with the methodology of the support, leading to more trust and more influence of the advice. A closer inspection of the rationales of the participants confirms this explanation. For instance, one participant interpreted "characteristics of the authors" as "what the authors have written before" and thus found that a very good indicator. However, a different participant interpreted the same phrase very differently as "how long this author has been active". The observation that participants created their own interpretations may have caused the high trust in the support, as well as the high ratings on other aspects such as usefulness and perceived speed of judgment.

In contrast to the complex support, the simple support was very clear about the features on which the advice was based (i.e., number of words, references, and images). This means that one of the main preconditions for a positivity bias (little information about the object) was not met. Hence, it is likely that no such bias occurred. This means that the participants' trust was based to a larger extent on the actual methodology of the system, of which they seemed to disapprove. This led to low trust in this simple support, and only very limited influence of this system on trust in the information. The observation that the features employed by the simple support were inappropriate was quite unexpected, since these features were directly derived from cues found to be used by a similar group of participants (Lucassen et al., 2013; Lucassen & Schraagen, 2010). Still, in this experiment numerous participants complained that these indicators were overly simplistic. In their motivations, the participants often pointed out that the heuristics were not always valid, for example "a short article does not necessarily mean that its contents are inaccurate" or "this article has a lot of references and pictures, but are they also high-quality?" The observation that the participants did not accept the validity of the decision rules based on features mentioned by a similar group of participants may indicate that although these features are somehow incorporated in credibility evaluation, this is not in a linear fashion as done by the support. It was for instance suggested by Lucassen and Schraagen (2010) that some features (for instance text length and number of references) are used as indicators of low credibility, but not as indicators of high credibility. In other words, a short text may indicate low credibility, but a long text does not necessarily indicate high credibility; the relationship between the quantity of the feature and its impact on credibility is thus not necessarily linear. Moreover, it may have been the case that simple decision rules were applied by the participants in our previous research, but only implicitly, as they had to think aloud while performing their task. In the current experiment, the decision rules were made explicit, which may have led the participants to realize that they were too simplistic. Dissociations between concurrent and retrospective reports have been

reported before (Ericsson & Simon, 1984), as well as dissociations between verbal reports and actual or predicted performance (e.g., Broadbent, Fitzgerald, & Broadbent, 1986). Similarly, Cooke and Breedin (1994) found dissociations between individuals' written explanations for physics trajectory problems and their predictions of those trajectories. In credibility evaluation, comparable differences between reported behavior and actual performance have been shown. Walraven, Brand-Gruwel, and Boshuizen (2009) found that high-school students could report numerous aspects on which they based credibility, but did not actually apply them in real-life.

In conclusion, we find it likely that a large part of the success of the complex support system was due to a positivity bias towards the support, which was absent in the simple support system. This implies that there is a risk over inappropriate reliance (Lee & See, 2004). Hence, when developing complexer support systems, this risk should definitely be taken into account. A countermeasure against over-reliance is to provide the user with enough information about the support to have the opportunity to assess it appropriately, in line with the conceptual model of the dynamics of trust (Lee & See, 2004). This can be done by providing sufficient information about how the system works. In case this is too complex for the average user, one could also pinpoint situations in which the support may perform less well (Dzindolet et al., 2003).

Finally, it should be noted that while the complex support system scored higher on most of the measures, confidence was boosted by the simple support, and only to a lesser extent by the complex support. The reason for this motivation is as yet unclear. One potential explanation lies in the formulation of the question on confidence. Participants were asked how confident they felt about their trust judgment. It may be that because they disagreed with the methodology of the support, they felt even more confident about their own judgment. For the other systems (complex automated and user-based), they agreed with the methodology to a much larger extent. Especially in cases where they would have rated the article differently than the advice from the support, they may have felt less confident about their final judgment (i.e., the combination of their own impression and the advice of the support). If this explanation is right, the boosted confidence is in fact not an argument in favor of the simple support system. Rather, it demonstrates that although users followed the advice of the other support system(s), they had some reservations about the advice.

4.2 User-based vs. Automated Support

Since the complex support seemed to perform better in this experiment, we compare that system to the user-based reputation system. The results showed that whereas the complex automated support could influence trust in the information both positively and negatively, the user-based reputation system could only diminish trust when giving negative advice. However, the impact of negative advice on trust in the information was larger than for the automated support. Furthermore, they were both similarly trusted, confidence levels were the same, as were the opinions on the support.

As argued before, the influence of negative advice on trust in information is perhaps the most important performance indicator of support systems, as this can lead users away from less credible information. In this study, it was observed that the user-based reputation system did this to a larger extent than the automated support systems. The tendency to attribute more value to negative user reviews than positive reviews (i.e., a negativity bias) has been shown before by Sen and Lerman (2007). They showed that for utilitarian products, negative reviews were trusted more than positive reviews. However, this was not the case for hedonic products. User reviews are very similar to a user-based reputation system, apart from the level of information they provide (a written review instead of only a numerical rating). A potential explanation for the difference between user-based and automated support in this study can be found in the reason why the participant had trust in the support. As discussed in the preceding section, trust in the complex automated support was possibly caused by a positivity bias because of a lack of information about the support. However, the user-based support is likely to be much more familiar to the participants, as many websites nowadays feature them. Therefore, it is likely that trust in the support came from earlier experiences which such systems (similar to history-based trust, cf. Merritt & Ilgen, 2008) rather than a mere positivity bias. This type of trust is likely to be stronger than the dispositional trust influenced by a positivity bias. Considering the high base level of trust in the articles (i.e., trust without support), negative advice from the support may have seemed in stride with their own judgment. Therefore, the support needed to be rather persuasive in order to be able to diminish the participants' trust. The user-based reputation system was more successful in doing so, likely due to the stronger type of trust in the support.

However, the experience that the participants already had with user-based reputation systems also led them to identify two practical, but critical shortcomings of DSS in this situation, namely the limited number of votes and the lack of information about the voters.

In this study, we simulated a limited number of votes (around 90) to replicate a potential drawback of the application of user-based reputation systems in the domain of collaboratively generated content, namely the dynamism of the information (Korsgaard & Jensen, 2009). On websites such as Wikipedia, information tends to change regularly over time, as contributions are immediately published on the website. This means that one of the requirements of Resnick et al. (2000) was violated in this case, namely that the objects of interest should be long-lived. Strictly, all votes on earlier versions of articles should be dismissed as soon as an article is updated, because 1) the edit could have resolved issues which caused a low rating or 2) the edit could have created new issues, rendering the existing rating to be too high. The number of votes in this study was higher than can be observed on Wikipedia, but still many participants stated that it was too low.

Moreover, many participants complained about the credibility of voters, which they found doubtful. How do we know that when someone voted, he or she has sufficient knowledge about the topic to know whether the information is credible? The advice given by the reputation system in this study did not provide any cues about the credibility of the voters. One could reason that this lack of information would again lead to a positivity bias, as it did with the complex automated support. However, for numerous participants this was clearly not the case, as reflected by their motivations. An explanation for the absence of a positivity bias can be found in the experience of the participants with reputation systems; they may have been aware of the fact that such systems are not always reliable.

A potential solution for this issue (and perhaps also for the limited number of votes) is to add the option for voters to write a short motivation for their opinion on the information. By adding this, subsequent users gain more cues to appropriately value the votes of preceding users. For instance, a very low rating without any motivation can then be valued as less important than a similar rating with a well-written motivation. Moreover, the threshold for malicious voters increases, as the required effort is slightly higher.

The limited influence of positive advice from any of the support systems on trust in the information may also be explained by the rather high level of trust that the information already received without support. The articles were intended to be of ambiguous quality; this was done by selecting articles from a rather low quality class on Wikipedia. Nevertheless,

this study showed that these articles were highly trusted, which may be an indication of high trust in Wikipedia, or the relatively high quality of these articles, despite their lower quality class. This high baseline trust may have resulted in a very limited influence of positive advice, as the support was unable to increase trust even further. Future research could include information of even lower quality to investigate this issue further.

4.3 Limitations

The results of this study should be considered with some caution. First of all, the participants in this study were university students who were aware of the fact that they were participating in an experiment on the effectiveness of supporting credibility evaluation. While students are regular Wikipedia users (Head & Eisenberg, 2010), they are not representative for the entire population of internet users. They are highly educated, with presumably good information skills. This may have consequences for their use of support systems. Moreover, they may have been more critical than the average user in the judgment of the various support systems. These observations mean that the conclusions may not generalize to the population at large. More research in this direction, featuring other user groups is highly recommended.

Second, the three support systems were evaluated in the context of Wikipedia. While this provides a great case study, some specific characteristics may be very different from other online sources. One particular difference is the open-editing model behind Wikipedia. For each of the conclusions and recommendations, the rationale behind it should be kept in mind to check whether it also applies in non-collaborative environments.

5. Conclusions

This study has demonstrated that several aspects should be dealt with when developing decision support for online credibility evaluation. Based on the experiment, we cannot recommend one particular support system, as each of them has its own drawbacks. However, we are able to identify several challenges one faces when developing support for this task. While each of the support systems had its own problems leading to less trust and thus less influence on trust in the information, most of the issues can be attributed to one thing, namely the level of detail of the information that the participants had. Only in the case of the simple automated support, the participants had all the information they needed to make a well-grounded claim about the credibility of the support. In this case,

they did not agree with the methodology, so the simple automated support was largely discarded. The complex support system was much more vague in its description. This possibly led to a positivity bias towards the support, and the observation that participants filled in the blanks about the inner workings of the system themselves, also leading to more trust. In the case of the user-based reputation system, a lack of information about the credibility of the voters led to less trust.

It can thus be concluded that the level of information provided about credibility evaluation support is a vital consideration. This may pose a serious challenge in cases where the support is too complex for the user to fully understand. When opting for a user-based reputation system, the dynamism of the information needs to be kept in mind. Automated support systems do not suffer from this drawback, but they should work in a way that the user understands (for appropriate reliance) and agrees with (for trust).

Chapter 8

The Role of Topic Familiarity in Online Credibility Evaluation Support

Evaluating the credibility of information is a difficult yet essential task in a society where the Internet plays a large role. Familiarity with the topic at hand has been shown to have a large influence on the way credibility is evaluated; 'familiar users' tend to focus on semantic features, 'unfamiliar users' focus more on surface features. In this study, we attempt to find out whether these differences have consequences for the development of credibility evaluation support systems. Two simulated support systems were evaluated, one utilizing semantic features (aiming at familiar users), the other utilizing surface features (aiming at unfamiliar users). The results suggest that unfamiliar users have a preference for a surface support system. Familiar users have no clear preference, have less trust in the support, and report to be influenced less by a support system. We recommend to focus on unfamiliar users when developing credibility evaluation support.

Published as Lucassen, T., & Schraagen, J. M. (2012). In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting (HFES 2012), October 22-26, 2012.*

1. Introduction

Nowadays, Internet users can easily obtain all information imaginable. However, not all information found online is equally credible. Therefore, users should always keep in mind that information retrieved may contain errors. Evaluating the credibility of information is a difficult task, for which most users are not specifically trained (Hargittai, Fullerton, Menchen-Trevino, & Thomas, 2010). This may lead to the problem of inappropriate trust (Hoffman et al., 2009), as users do not know how to evaluate credibility.

One approach to help Internet users to evaluate credibility is to offer them decision support tools (Pirolli, Wollny, & Suh, 2009). Such tools can give advice about the credibility of information, which can be used to form an appropriate level of trust, combined with the user's own evaluation. Various attempts have been made to support users in their evaluation, such as reputation systems (Jøsang, Ismailm & Boyd, 2007) or automated tools (e.g., Yamamoto & Tanaka, 2011). Currently, automated attempts focus primarily on finding optimal cues in the information, which can be measured automatically, and correlate well with information quality. In such approaches, the human factors aspect of credibility evaluation support is often neglected, potentially leading to poor user acceptance (Ye & Johnson, 1995).

Research on trust in online information has shown that the strategies applied to evaluate credibility are largely dependent on various characteristics of the user (Lucassen & Schraagen, 2011b). Users with a certain level of domain expertise (or knowledge) can bring their knowledge to bear when evaluating information by comparing the information presented with their stored knowledge of the topic (Eastin, 2001). Domain novices are less capable of making such comparisons, but can work around this limitation by applying their information skills (Lucassen, Muilwijk, Noordzij, & Schraagen, 2013). By doing so, they do not consider the factual accuracy of the information, but the manner of presentation (e.g., length of the text, number of references). This result is highly similar to the well-known semantic-surface distinction in expert-novice studies (e.g., Chi Feltovich, & Glaser, 1981).

Hence, different users evaluate the credibility of information differently. However, existing support systems do not account for such differences, because they aim at technological sophistication rather than usability. In this paper, we consider the implications of one

particularly important characteristic of the user for credibility evaluation support, namely topic familiarity.

Familiarity can be seen as a weak form of domain expertise. It is likely to vary greatly between various information searches, as the topics Internet users search for can both be close to their own expertise (e.g., profession), or more distant (e.g., searching for medical care). Lucassen and Schraagen (2011b) showed that trust of experts in the domain of automotive engineering was influenced by factual errors in Wikipedia articles on this topic, whereas trust of novices was not. This difference was explained by the cues in the information used by the novices and experts; novices mainly focused on surface cues (presentation), whereas experts assessed both semantic (content) and surface cues.

Given the important role of topic familiarity in credibility evaluation, accounting for this factor when developing credibility evaluation support may be important for user acceptance. Each user has his or her own mental model of credibility (Hilligoss & Rieh, 2008). In order to maximize trust in the support itself, the method applied by the support (i.e., the cues it uses) should be approved by the user. It is more likely that this is the case when there is a match between the mental model of the user and the support model (Ye & Johnson, 1995; Cohen, Parasuraman, & Freeman, 1998).

In credibility evaluation support, problems with mismatches between the features utilized by a support system and the mental model of the user have been shown by Lucassen and Schraagen (2011a). In an evaluation of the WikiTrust system (Adler et al., 2008), even highly educated, young users had difficulties interpreting the advice of the support. This can possibly be attributed to the feature utilized by the support, namely edit age (parts of information which are older, are more credible). It is unlikely that the participants had a concept similar to the age of information in their mental model of trust, as similar users have reported very different cues in another experiment (Lucassen & Schraagen, 2010). Problems with edit age as an indicator for credibility have also been shown by Luyt, Aaron, Thian, and Hong (2008).

In this paper, we simulate two credibility evaluation support systems in the context of Wikipedia. The first of these systems uses *surface* cues to offer advice to the user; the latter uses *semantic* cues. We hypothesize that when evaluating information on a familiar topic, users prefer a support system that also incorporates semantic cues. In contrast, when evaluating information on an unfamiliar topic, they prefer a "surface system". We

tested these hypotheses through a lab experiment, in which the participants had to rate the credibility of Wikipedia articles with such support systems present. Topic familiarity was manipulated between subjects. After having experienced both systems on a few occasions, they had to choose one of them for the final trial. We take this choice as a main indicator of preference for one of the support systems.

2. Method

2.1 Participants

A total of 40 participants (9 male, 31 female) volunteered in the experiment. Their average age was 22.3 (*SD* = 2.0). 28 participants were German, the rest was Dutch. All participants were proficient in the English as well as the Dutch language. All participants were psychology students at a Dutch university. While this was above all a convenience sample, college students are frequent Internet and Wikipedia users (Head & Eisenberg, 2010), with a persisting need for new information. The participants reported daily (7.5%), weekly (57.5%), or monthly (35.0%) use of Wikipedia. Informed consent was signed by each of the participants prior to the experiment.

2.2 Task and Design

The participants were asked to perform the Wikipedia Screening Task (Lucassen & Schraagen, 2010). In this task, a Wikipedia article is displayed to the participant, who is then asked to evaluate its credibility. The way to perform this task is not specified, so the participants are free to choose and develop their own methods.

A 2 (familiarity) × 2 (available support system) × 2 (positive/negative advice) design was applied. Familiarity was manipulated between subjects; all participants received both positive and negative advice from both support systems.

Table 1: Articles used in the experiment.

Familiar topics	Unfamiliar topics
Abstinence	Chinoiserie
Austin Powers in Goldmember	Bob Black
Online community	Ruben's tube
Mental model	Richard Evan Schultes
Tamagotchi	Esoteric Cosmology

Figure 1: The simulated "Dispute Finder" support system giving positive and negative advice.

Figure 2: The simulated "WikiChecker" support system giving positive and negative advice.

Half of the participants evaluated articles on familiar topics (from now on called familiar users); the other half evaluated unfamiliar topics (unfamiliar users, see Table 1). Familiarity was manipulated by selecting topics on Wikipedia which are generally familiar or unfamiliar to Dutch and German college students. The manipulation was verified by letting the participants rate their familiarity with each topic on a 7-point Likert scale.

All articles were obtained from the English Wikipedia, as a quality assessment is available for most articles in this language (Wikipedia: Version 1.0 editorial team, n.d.). The articles used were all of "Start-class" quality, which can be denoted as "developing" articles. This implies that the participants had reasons to trust and distrust the articles. In other words: the credibility of the articles was ambiguous. At the top of each article (just beneath the article title), advice from one of the two simulated support systems was displayed. This advice was either positive or negative (counterbalanced between subjects).

The first support system used semantic cues to generate advice. This was done following the principle proposed by Ennals, Byler, Agosta, and Rosario (2010). Their "Dispute Finder" tool takes claims from Wikipedia and searches the Web for sentences falsifying these claims. The instructions for the participants stated the following:

> *The Dispute Finder searches for websites disputing the information in the Wikipedia article. This is done by searching for patterns such as "falsely claimed that …" or "the misconception that …". Based on the number of "disputed claims" found, the Dispute Finder gives positive or negative advice about the credibility of the article.*

In our simulation, the advice simply stated that it could or could not find contradicting statements on the Web. The Dispute Finder is shown in Figure 1.

The second support system, shown in Figure 2, used surface (non-semantic) cues (e.g., details of the edit history and reference list) to generate advice. The system, labeled

"WikiCheck", was introduced as a complicated computational system in the instructions for the participants:

> *The WikiChecker uses an adaptive neural network which gives a credibility rating based on numerous features of the article (e.g., author details, edit history, reference list). Based on this algorithm, the WikiChecker gives positive or negative advice about the credibility of the article.*

The presented articles with the given advice were actually full-page screenshots of the actual Wikipedia articles. This was done to keep the participants from following the links in the article and the advice.

2.3 Procedure

Upon arrival, the participants and experiment leader first signed a consent form. Aside from this form, the entire experiment was conducted digitally through the "LimeSurvey" online questionnaire tool (Schmitz, 2012).

On the first screen, the participants were instructed on the course of the experiment. After clicking "Next", the participants had to fill in their age, gender, nationality, study progress, and Wikipedia use. Moreover, they were asked to explain in their own words how Wikipedia works. All participants mentioned the open editing model behind Wikipedia. Finally, the participants were asked how credible they thought information on Wikipedia normally is.

On the following screen, all five topics to be evaluated in the experiment were listed. The participants were asked to rate their familiarity with these topics on a 7-point Likert scale. These scores were compared afterwards to check the manipulation of familiarity.

Subsequently, the actual experiment started. First, an explanation of the first support system (the Dispute Finder) to be used was given. After this, two articles with this support system were presented to the participants on separate pages. Below each article, the participants were asked about:

- Trust in the article (along with a motivation)
- Trust in the support
- The perceived influence of the support on their trust in the article

These questions had to be answered on 7-point Likert scales. After evaluating two articles, the explanation of the second support system (WikiCheck) was given. After reading this, two articles were displayed on separate pages, along with the same questions as with the first support system.

After evaluating four articles in total, the participants were instructed that they would have to evaluate one more article, and that they were free to choose one of the two support systems. The final article featured the support system of their choice. Again, the same questions as with the first four articles had to be answered (but for balancing reasons, were not included in the analyses).

After finishing the experiment, the participants were informed about the goal of the experiment and explained that the support systems were actually simulated. The whole procedure took approximately 30 minutes.

All statistical analyses were performed using non-parametric tests, since Likert-scales are assumed to be measuring at the ordinal level rather than the nominal level (Jamieson, 2004).

3. Results

3.1 Manipulation checks

The manipulation of familiarity was verified by comparing the familiarity ratings of the presumed familiar articles with the unfamiliar articles. The manipulation proved successful as familiarity was rated higher for the familiar articles ($M = 3.50$, $SD = .96$) than for the unfamiliar articles, ($M = 1.23$, $SD = .37$) $U = 4.00$, $p < .001$.

3.2 Preference for support system

Table 2 shows the preferences on the final trial of familiar and unfamiliar participants for either of the two support systems.

Unfamiliar users showed a near-significant preference for the WikiChecker support system over the Dispute Finder, $\chi^2(1) = 3.20$, $p = .07$. Familiar users showed no significant preference for either of the support systems, $\chi^2(1) = 0.20$, $p = .66$. Overall, the participants had no significant preference, $\chi^2(1) = 0.90$, $p = .34$.

Table 2: Number of participants preferring each support system.

	Dispute Finder	WikiChecker
Familiar	11	9
Unfamiliar	6	14
All	17	23

Table 3: Average trust in the support on 7-point Likert scales. Standard deviations are given in parentheses.

	Dispute Finder	WikiChecker
Familiar	4.35 (1.18)	4.28 (0.94)
Unfamiliar	4.55 (1.21)	4.86 (0.90)
All	4.45 (1.19)	4.57 (0.96)

3.3 Trust in the support systems

Table 3 shows the level of trust in the Dispute Finder and WikiChecker support system in the familiar and unfamiliar condition.

Participants evaluating articles on unfamiliar topics had significantly more trust in the WikiChecker support system than in the Dispute Finder, $Z = 2.10$, $p < .05$. Participants evaluating familiar topics trusted both systems equally, $Z = 0.24$, $p = .81$.

Moreover, participants in the unfamiliar condition had more trust in the support systems than participants in the familiar condition, $U = 116.5$, $p < .05$.

3.4 Trust in the articles

Table 4 shows trust in the articles in the familiar and unfamiliar condition with positive and negative advice from both support systems.

Participants evaluating familiar topics had more trust in articles with positive advice than

Table 4: Trust in the articles when the support systems gave positive or negative advice on 7-point Likert scales. Standard deviations are given in parentheses.

	Dispute Finder		WikiChecker	
	Positive	Negative	Positive	Negative
Familiar	5.15 (1.50)	4.45 (1.10)	5.45 (1.00)	4.65 (1.42)
Unfamiliar	5.30 (0.92)	4.60 (1.00)	5.85 (0.59)	3.85 (1.27)
All	5.23 (1.23)	4.52 (1.04)	5.65 (0.83)	4.25 (1.39)

Table 5: Number of motivations for trust in the articles in each category. Categories indicate surface/semantic strategies, and following of the advice of the support.

	Familiar	Unfamiliar
Surface features	40	66
Semantic features	32	2
Advice of the support	17	30
Other	11	2

Table 6: Reported influence of both support systems on 7-point Likert scales. Standard deviations are given in parentheses.

	Dispute Finder	WikiChecker	All
Familiar	3.48 (1.44)	3.33 (1.63)	3.40 (1.33)
Unfamiliar	4.13 (1.15)	5.00 (0.95)	4.56 (0.83)
All	3.80 (1.32)	4.16 (1.57)	3.98 (1.24)

in articles with negative advice when given by the WikiChecker system ($Z = 2.15$, $p < .05$), but not by the Dispute Finder ($Z = 1.44$, $p = .15$). With unfamiliar topics, trust in the articles was higher with positive advice than negative advice regardless of the support systems (WikiChecker: $Z = 3.76$, $p < .001$, Dispute Finder: $Z = 2.13$, $p < .05$).

Table 5 shows the participants' motivations for trust in the articles. Classification was done in accordance with Lucassen and Schraagen (2011b). 25% of the answers were coded by two experimenters. The inter-rater reliability as expressed by Cohen's Kappa for this overlap was .90.

Motivations in the surface category largely regarded the references, structure, and style of the article. Motivations in the semantic category were mostly about accuracy, completeness, and neutrality. Finally, motivations regarding the advice of the support system largely stated that they followed the advice of the support.

Familiar users reported to have considered significantly different features than unfamiliar users, $\chi^2(3) = 42.67$, $p < .001$. The number of motivations in all categories listed in Table 5 differ significantly between groups.

3.5 Reported influence of the support

Table 6 shows the reported influence of either of the support systems on the credibility evaluation. Participants in the unfamiliar condition reported to be influenced by the

support more than participants in the familiar condition, $U = 81.5$, $p < .01$. In the familiar condition, participants reported to be similarly influenced by both support systems, $Z = 0.35$, $p = .73$. In the unfamiliar condition, the participants reported to be influenced more by the WikiChecker than the Dispute Finder, $Z = 2.55$, $p < .05$.

4. Discussion

In this study, we examined how users evaluating the credibility of familiar or unfamiliar topics differed in their preference for support. Two simulated support systems were tested, one utilizing semantic features (aiming at familiar users), the other utilizing surface features (aiming at unfamiliar users). The results showed that unfamiliar users:

- (marginally) preferred surface support;
- had more trust in surface support than semantic support;
- were influenced by both support systems;
- reported to be influenced more by surface support than by semantic support.

In contrast, familiar users:

- had no preference for either of the support systems;
- had less trust in both systems than unfamiliar users, with similar (low) trust for both support systems;
- were solely influenced by surface support;
- reported to be influenced similarly (low) by both support systems.

Based on these outcomes, the following conclusions can be drawn. First, familiar users were largely indifferent in their opinion (preference, trust, and reported influence) about the support systems. This can be explained by the fact that domain experts (of which familiar users are a weaker form) do not only consider semantic features in their evaluation, but also surface features (Lucassen & Schraagen, 2011b; Lucassen et al., 2013). Table 5 indeed shows that despite their familiarity, a large portion of the motivations were still at the surface level. Each of the two support systems only utilized one of the two categories of features (surface/semantic), which may mean that they have been perceived as too limited by the familiar users, who use both categories. This explanation is supported by the observation that familiar users had less trust in the support than unfamiliar users, and reported to be influenced less by the support.

Furthermore, Metzger (2007) proposed that ability of the user largely determines the extent to which an evaluation is performed. As the arsenal of strategies to evaluate is

larger for familiar users, they may have felt less need for an automated system to support them in their evaluations.

As expected, unfamiliar users showed a tendency towards the surface support system. This tendency can be ascribed to the utilization of similar features as the participants. Table 5 shows that about two-thirds of all motivations could be categorized as being at the surface level. An alternative explanation is that based on the instructions of the WikiChecker system, it was perceived as more advanced than the Dispute Finder. No specific evidence for this explanation could be found in the results, but future research could focus on a more strict manipulation of surface and semantic support, whilst sustaining the system instructions.

Trust of unfamiliar users was influenced by both the semantic and the surface support system. This is of course not a strong indication of their preference for the surface support. However, because of their limited ability to evaluate (i.e., no domain expertise), their need for support was presumably much larger than that of familiar users, which means that they may have taken any advice available to aid them, regardless of its methodology. Still, when given a choice they seemed to prefer the surface support.

5. Recommendations

Given these findings, we recommend to keep the familiarity of the users in mind in the development of credibility evaluation support systems. We recommend to focus primarily on unfamiliar users. These users are at the largest risk of trusting poor quality information, as their evaluation skills are generally more limited than those of familiar users (i.e., lack of domain expertise).

The recommendation to focus on unfamiliar users when developing credibility evaluation support, does not mean that only surface features should be used in such systems. However, the mental model of trust of the user should be kept in mind. When the users approve of the methodology applied by a support system, acceptance is likely to be higher (Ye & Johnson, 1995) and trust in the support more appropriate (Hoffman et al., 2009).

Acknowledgments

The authors would like to thank Lotta Schulze for her efforts in gathering the data in the experiment.

Chapter 9

Summary & Conclusions

1. Summary

In the introduction of this dissertation, I stated that our current knowledge of trust in online information lacked understanding of how user characteristics influence the process of credibility evaluation and, ultimately, trust. Several theories and models have touched upon this topic (e.g., Fogg, 2003; Hilligoss & Rieh, 2008; Metzger, 2007), but none of them are very specific about how the interaction between user characteristics and information features works.

In Chapter 2, we started to address this knowledge gap by identifying three user characteristics that we hypothesized to be particularly influential for credibility evaluation, namely domain expertise, information skills, and source experience. We brought these together in a novel model, named the 3S-model of information trust. In this model, we proposed that domain expertise has a direct influence on the usage of semantic features (1st S), information skills on the usage of surface features (2nd S), and source experience on the usage of source features (3rd S). A combination of the evaluation of features in each of these categories will lead to a judgment of trust in a piece of information. Initial validation for the 3S-model was provided by letting self-selected experts and novices in the domain of automotive engineering evaluate the credibility of Wikipedia articles on car engines. The factual accuracy (a semantic feature) of these articles was manipulated to different degrees, with up to 50% factual inaccuracies. The validity of the semantic part of the model was demonstrated by the fact that experts were influenced by the varying accuracy, whereas novices were not. However, we also found that experts were only influenced to a limited extent by the accuracy of the information. This was regarded as evidence that, apart from semantic features, experts also apply information skills and source experience during credibility evaluation. This notion was supported by a categorization of the motivations provided by the participants, indicating that both experts and novices use all three strategies in their evaluations.

The primary aim of Chapter 3 was to provide additional validation for the proposed 3S-model, with a primary focus on information skills. This was achieved through an experiment in which three different groups of students (high-school students, undergraduate students, and postgraduate (PhD) students) were asked to evaluate the credibility of Wikipedia articles while thinking aloud. These groups were assumed to differ in their level of information skills. Articles were selected on an individual basis, after

asking each of the participants in a telephone interview prior to the actual experiment about topics they were particularly familiar and unfamiliar with. The results largely supported the proposed model. Higher familiarity (a weaker form of domain expertise) led to more frequent use of semantic features for all groups. Moreover, students with better information skills (i.e., undergraduates, postgraduates) used more surface features than hose with poorer information skills (i.e., high-school students). In contrast to Chapter 2, no effect of familiarity on trust was found. This could be explained by the different manipulation of quality in the experiment in Chapter 2 and in this experiment. In Chapter 2, factual accuracy was manipulated by the experimenter, whereas in this experiment, the manipulation followed the ratings of the Wikipedia Editorial Team. It can be expected that quality differences following these ratings are predominantly visible at the *surface* level, since they represent how far an article is from a distribution-quality encyclopedic article. Factual accuracy (in Chapter 2) is a *semantic* feature. Hence, in Chapter 2, familiarity influenced trust because a semantic feature was manipulated, and in Chapter 3, information skills influenced trust because the manipulation was mainly visible at the surface level. This explanation was supported by the observation that only trust of students with better information skills was influenced by the quality of the information.

The use of one particular surface feature was studied in Chapter 4, namely the references at the bottom of Wikipedia articles. This was done because undergraduate students demonstrated a particular interest in this feature in Chapter 3. In the experiment, references of Wikipedia articles were manipulated on two dimensions, namely quantity and quality. The first was manipulated by varying the number of references (from about 5 to about 25); the latter by replacing all references of a Wikipedia article by those of a different, completely unrelated article. We expected that the quantity manipulation could be detected easier than the quality manipulation. The first would require only some heuristic processing of this feature, whereas the latter would require more rigorous, systematic processing. The results showed that both strategies (heuristic and systematic) were applied by the participating students. Interestingly, the choice between the strategies was shown to largely depend on trust in the source (Wikipedia) in general. Participants with a skeptical attitude towards Wikipedia tended more towards systematic evaluation, whereas those with a less critical attitude only performed a heuristic evaluation. Moreover, despite the great interest for references demonstrated in Chapter 3, only 26% of the participants in this experiment had actually noticed the quality manipulation. For the quantity manipulation, the percentage was somewhat higher with 39%. We coined this

phenomenon "Reference Blindness": references are an important indicator for students, but as long as they are present, their quality (and, to a lesser extent, quantity) hardly seems to matter.

The third 'S' of the 3S-model, source experience, was the main topic of research in Chapter 5, together with domain expertise. In an earlier study (Eastin, 2001), it was hypothesized that the source of information was less important in credibility evaluation when the user was familiar with the topic at hand because in that case the user could focus on the actual content (semantics) of the information. However, in that earlier study, no conclusive support for this could be found. In the experiment described in Chapter 5, we investigated this issue further, by showing participants Wikipedia articles in their original layout (thus with the presence of source cues), or in a standardized layout (based on Microsoft Word 2003, thus without the presence of source cues). In contrast to the original hypothesis of Eastin (2001), trust of unfamiliar users was not influenced by the presence of source cues. Familiar users were influenced, but only when the quality of the information was questionable. Under these circumstances, familiar users had less trust in the information when they knew it came from Wikipedia. The difference between our results and Eastin's hypothesis can be explained by the open-source character of the source used in this study, namely Wikipedia. Familiar users may value their knowledge on the topic at hand over the knowledge of other (unknown) people.

The three user characteristics proposed in the 3S-model were placed in a broader perspective in Chapter 6. In this chapter, we proposed a new research model, in which multiple layers explain the influence of various user characteristics on trust. The core of this layer model is trust in the information itself (based on evaluation of semantic and surface features). The first layer around this core represents trust in a source. The second layer is trust in a medium (a generalization of trust in multiple, comparable sources; e.g., the Internet, newspapers), and the outer layer concerns a general propensity to trust. We hypothesized that each layer has a direct influence on the next, adjacent, layer. Its influence on distant (non-adjacent) layers is mediated though the intermediate layer(s). Bootstrapping mediation analysis supported all of the hypothesized relationships. Post-hoc analyses also demonstrated that users with little trust in the source (in this case, Wikipedia) did not differentiate between high-quality and low-quality information, whereas users with more trust in the source did. This indicates a negative biasing effect of the source of information for distrusting users, but no positivity bias for trusting users.

The 3S-model and other theories on trust in online information show numerous potential problems that Internet users have with evaluating credibility. Examples are a lack of domain expertise or information skills (3S-model), and motivation or ability (Metzger, 2007). Therefore, we explored the application and user acceptance of decision support systems in the domain of online information in Chapter 7. Three simulated support systems were offered to participants in a lab experiment, thereby considering two important trade-offs, namely the choice between user-based or automated support, and in case the latter is preferred, the choice between a complex (high performance, low comprehensibility) and simple (low performance, high comprehensibility) system. Concerning the choice between user-based or automated support, no definitive conclusions could be drawn based on the results. Both have advantages and disadvantages. Trust in user-based support is heavily dependent on the number and credibility of the users; automated support is susceptible to inappropriate reliance, as its reliability is not likely to be 100% in complex, dynamic environments as the Internet. Simple automated support was discarded quickly by the participants, as they regarded the principles on which is was based far too simplistic to be effective.

In Chapter 8, a direct link between the theory behind the 3S-model and the application of automated support systems was made. Various studies (e.g., Chapters 2 and 3) have demonstrated that domain experts (or familiar users) evaluate credibility differently than domain novices (or unfamiliar users). The key difference is the utilization of semantic features, which is done to a much greater extent by domain experts. In Chapter 8, we investigated whether this has consequences for the application of support systems. Participants familiar and unfamiliar with certain topics were offered two types of support system. The first system based its advice on semantic features (typical for domain experts), whereas the second system based its advice on surface features (typical for domain novices). After having experienced both systems, the participants could choose one of them for the final trial. The choice they made was seen as an indication for their preference. Unfamiliar users showed a clear preference for the support that used surface features, whereas familiar users showed no preference. The other measures in the experiment (e.g., trust in the support, reported influence of the support) showed a similar pattern. We concluded that when developing decision support systems for credibility evaluation, it is best to focus on unfamiliar users, as familiar users are better equipped to evaluate the credibility of the information themselves.

2. Conclusions

The various chapters of this dissertation contributed to answering the original research question, namely the influence of user characteristics on credibility evaluation. Many ways in which various user characteristics influence credibility evaluation, trust, and each other were demonstrated, and the relationships between these constructs were captured in the 3S-model. The layer model introduced in Chapter 6 placed the studied characteristics in a broader perspective. Figure 1 shows an integrated model, based on the two models introduced in this dissertation. In this model, I integrate the empirical findings of the experiments reported in this dissertation.

The model depicted in Figure 1 concerns the behavior of an Internet user who encounters a new piece of information. First, the user decides (implicitly or explicitly) whether there is a need to evaluate the credibility of the information. This need may for instance stem from the consequences of poor information (i.e., motivation; Metzger, 2007), or the perception of risk (Corritore, Kracker, & Wiedenbeck, 2003). Another reason for a user to feel the need to evaluate the information may originate from prior experiences with the source of the information. For instance, when a person has bad experiences with a source, the need for a stringent evaluation is high if he or she wants to use the information (Chapter 4). Bad experiences may also lead a user away from a credibility evaluation, as he or she may disregard the information whatsoever (Chapter 6). The difference between these findings can be explained by the experimental setting: the experiment in Chapter 4 was a controlled lab experiment, whereas Chapter 6 described an online quasi-experiment, with much less control over the participants. The participants in the lab experiment may have felt more need to evaluate the information because they were asked to do so, even though they did not trust the source. For the participants in the online experiment, it was easier to just ignore the articles.

We demonstrated in Chapter 6 that trust in a source is largely influenced by trust in the medium through which this source is communicated. Trust in a medium in turn is influenced by a general propensity to trust. The influence of these factors on trust, even before the actual information itself is considered, has been suggested before in literature (Corritore et al., 2003; Hilligoss & Rieh, 2008).

When a user decides not to evaluate credibility, trust in the information is thus solely based on trust in the source of information. In dynamic environments such as the Internet, this

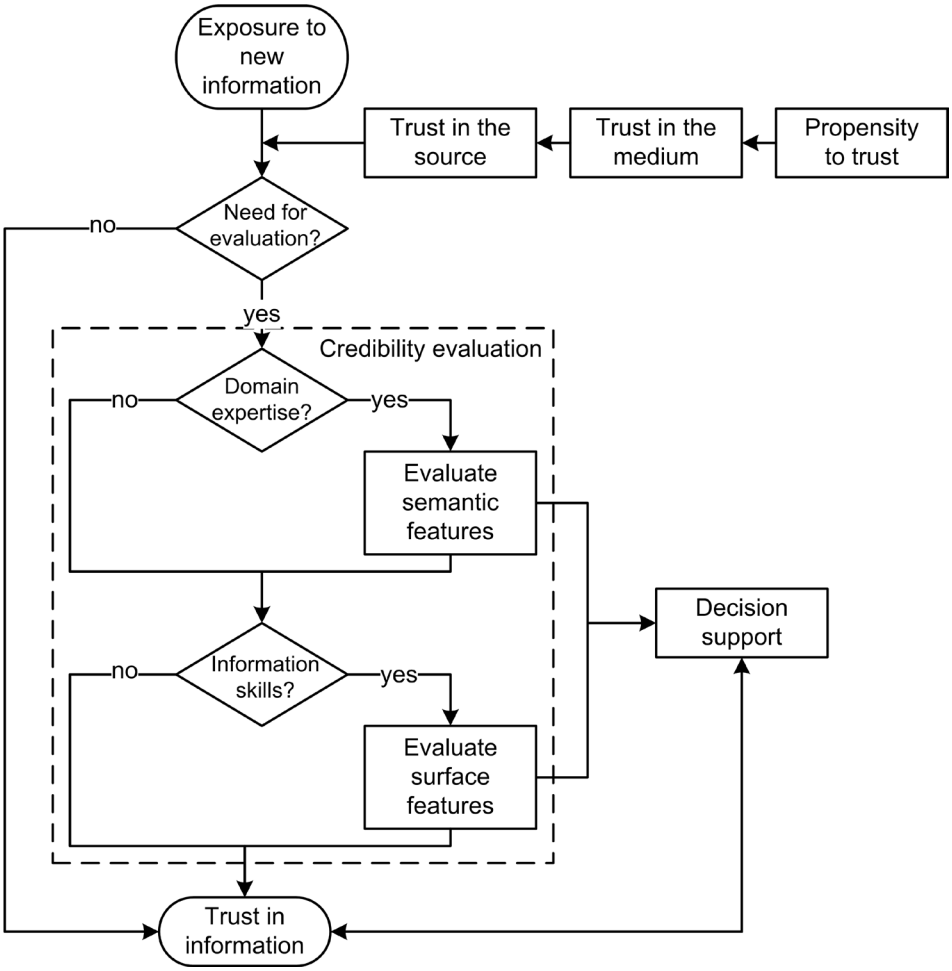

Figure 1: Integrated model of trust in online information.

may mean that the level of trust in the information is poorly calibrated with the actual quality of the information, as the source of information has only limited predictive power for credibility (Sundar, 2008).

When a user decides to perform a credibility evaluation, the calibration of trust can be improved. How such an evaluation is actually performed, depends on the domain expertise and information skills of this person. When the user is familiar (a weaker form of domain expertise) with the topic at hand, he or she is able to evaluate semantic features of the information (Chapters 2 and 3). When the user has a sufficient level of information skills, he or she can put these skills to use by evaluating surface features of the information (Chapter 3). It can thus be inferred that trust of users with domain expertise and information skills is likely to be best calibrated with the actual information quality.

In other words, users who know a lot about a particular topic and who at the same time possess a lot of sophisticated information skills, are best equipped to assess high-quality information in terms of credibility. Users without any of these characteristics may attempt to evaluate credibility, but it is unlikely that their trust is much better calibrated than when no evaluation is performed at all (cf. the high-school students in Chapter 3).

Different users thus evaluate the credibility of information differently. This is also predicted by the unifying framework of credibility assessment by Hilligoss and Rieh (2008). In the construct level of their framework, they posited that the personal definition of trust or credibility of a user is very important for the way that credibility is evaluated. We extend this notion with our view that various user characteristics lead to different features being evaluated; we propose that user characteristics such as domain expertise and information skills to a large extent influence the personal definition of credibility. Domain experts are likely to interpret credibility as "factual accuracy", as they tend to focus on semantic features. Users with high information skills are more likely to consider information credible when it is, for instance, well-supported by references.

A similar explanation for the differences in credibility evaluation between users can be found in the Prominence-Interpretation Theory by Fogg (2003). Following this theory, an element of a website can impact credibility assessment under two conditions. First, the element needs to be prominent for the user, and second, it needs to be interpreted. Both these aspects are affected by several personal and contextual factors. I contribute to this theory by proposing that domain expertise and information skills of the user can be interpreted as personal factors affecting prominence and interpretation. For domain experts, semantic features will be prominent, and for users with high information skills, surface features will be prominent. Moreover, domain experts will be able to attribute value to semantic features (interpretation), whereas users with high information skills can attribute value to surface features.

The MAIN-model of Sundar (2008) also focuses on the cues that signal credibility of information, which are conveyed by the affordances that Internet technology has yielded. He has identified numerous heuristics based on these cues. However, it is unlikely that users apply all of these heuristics. By applying the model depicted in Figure 1 we can predict which of the heuristics are likely to be applied by users with certain characteristics. This means that when considering the heuristics that are used in a certain context, variables on two levels need to be considered, namely 1) the heuristics that are relevant for a certain

type of website (e.g., authority heuristics for encyclopedic information, bandwagon heuristics for social networks; Sundar, 2008, p. 91), and 2) the heuristics that are likely to be used based on the characteristics of the user (cf. Figure 1).

A distinction that is commonly made in studies on credibility evaluation is between heuristic (peripheral) versus systematic (central) processing (Chaiken, 1980; Chaiken & Maheswaran, 1994). According to the dual processing model of web site credibility assessment (Metzger, 2007), the degree to which online information is evaluated depends on the ability of the user, which "may be linked to users' knowledge about how to evaluate online information" (p. 2087). However, I question the validity of this claim. In terms of the integrated model introduced here, ability is defined as domain expertise and information skills. We stated that domain expertise leads to the use of semantic features in credibility evaluation, and that information skills lead to surface features. However, we also argued that both heuristic and systematic processing is possible in either of these feature categories (Chapters 2 and 3). Hence, *ability* (at least in terms of domain expertise and information skills) does not lead to heuristic or systematic processing, but to the use of semantic or surface features. Both modes of processing are possible in these feature categories. My claim is supported by the results of the experiment in Chapter 4, in which we demonstrated that a group with very similar information skills (college students) processed the references of Wikipedia-articles (a surface feature) both heuristically and systematically, based on their disposition towards the source. Moreover, comparing presented facts (a semantic feature) with one's own knowledge is considered heuristic processing (Klein, Calderwood, & Clinton-Cirocco, 1986), whereas verifying the validity of a claim just outside one's area of expertise may involve highly systematic processing (Kahneman & Klein, 2009). Hence, the user characteristics underlying the concept of *ability* studied in this dissertation do not directly lead to heuristic or systematic processing, but rather to the use of different information features.

Rather than high user abilities leading to systematic processing, I would argue that they lead towards better heuristics (i.e., being better predictors of information quality). Consider, for instance, the differences in credibility evaluation between high-school students, undergraduates, and postgraduates (Chapter 3). These groups, which are likely to differ largely in their information skills, all demonstrated the use of certain heuristics in their evaluations. However, whereas the heuristic "this does not conflict with my own knowledge on the topic, so I trust it" may suffice for high-school students, undergraduates came up with heuristics such as "It is supported by many references, so I trust it".

Postgraduates may even go beyond this heuristic, by applying heuristics such as "It is supported by references in good journals, so I trust it". Clearly, this demonstrates that users of varying information skills all apply heuristics in their evaluations, but that the heuristics of people with better information skills may be more predictive of the actual quality of the information.

The notion that *ability* does not affect the level of processing does not mean that the distinction between heuristic and systematic processing is not valid in the domain of credibility evaluation at all. Chaiken and Maheswaran (1994), for instance, demonstrated that high task importance (which can be seen as a form of motivation, or the need to evaluate credibility) can lead to systematic processing; low task importance always leads to heuristic processing. Still, credibility evaluation will always be a heuristic process to a certain extent. As I already argued in the introduction: credibility evaluation is reducing the risk of trusting information. If credibility was evaluated entirely systematically, there would not be any risk anymore, which renders the concept of trust obsolete, as this is replaced by certainty. This notion is supported by various studies in which credibility evaluation is exclusively discussed in terms of the application of heuristics (Metzger, Flanagin, & Medders, 2010; Sundar, 2008; Sundar, Knobloch-Westerwick, & Hastall, 2007).

The model depicted in Figure 1 thus adds much more detail to existing models as to how user characteristics actually influence credibility evaluation, and ultimately trust. The knowledge gathered on how Internet users evaluate credibility can also be put to use in the design of successful decision support systems in this domain. In Chapters 7 and 8, we explored the application of decision support in credibility evaluation. Following the model in Figure 1, this supplies the user with an extra source of cues about the credibility of information. Users can combine these cues with the cues that they found in the information themselves (Chapter 7). Decision support systems may base their advice on semantic or surface features from the information (Chapters 7 and 8). A different approach is to base the advice on the opinion of prior users (i.e., a reputation system, Chapter 7), hence the bi-directional relationship between trust in the information and the decision support system.

We demonstrated that some of the existing challenges in the development of decision support for credibility evaluation are also highly valid in this domain. First of all, we showed that it is likely that the first of the requirements for online reputation systems

listed by Resnick et al. (2000), namely a long life-expectancy of the object of trust, is likely to be violated in dynamic environments such as Wikipedia. Articles are updated regularly, which means that user opinions on their credibility have to be dismissed rather quickly. This is not only the case for Wikipedia, but also in other environments with user-generated content. This may have severe consequences for the application of reputation systems in such domains, as it is very hard to build up a reliable reputation.

Moreover, in the well-known trade-off between *trustable* and *trustworthy* support systems (Lee & See, 2004), we demonstrated that the latter (a rather simplistic, but understandable system) is not likely to be successful in the domain of online information. Participating students largely ignored this system as they found it too simplistic to be useful for a complex task setting such as credibility evaluation. It could thus be argued that the development of support for credibility evaluation should aim at more complex, but thus less understandable systems. However, this introduces the problem of appropriate reliance (Parasuraman & Riley, 1997). Hence, future studies should particularly focus on tackling this issue.

It can be concluded that, overall, the introduced models in this dissertation provide more detail to existing models and theories. A revision of an existing model is warranted as far as the 'ability'-factor in the dual processing model of web site credibility assessment (Metzger, 2007) is concerned; I conclude that higher ability does not necessarily lead to more systematic levels of processing. All in all, four main conclusions can be drawn:

- Trust in information is at first primarily based on trust in the source, which in turn is influenced by trust in the medium and a general propensity to trust.
- Users can improve the calibration of their level of trust with the information quality by performing a credibility evaluation.
- User characteristics have a large influence on whether and how credibility evaluation is performed.
- It is vital for user acceptance to keep user characteristics in mind in the development of decision support for credibility evaluation.

3. Implications

The studies discussed in this dissertation have implications for research in the domain of online information, in which the introduced integrated model, as well as the 3S-model

(Chapters 2 and 3) and the layer model (Chapter 6) can serve as a framework for future studies. Moreover, one can now make informed decisions on how to deal with the user characteristics discussed here in future experiments. Given the large influence of these characteristics, future experiments should control or manipulate them. We have also demonstrated that a different line of research, namely that of decision support systems, can benefit as well from the knowledge about how user characteristics influence credibility evaluation and trust (see, for instance, Chapter 8). Up until now, the development of support systems for credibility evaluation has largely focused on technological advances rather than user acceptance.

The models introduced in this dissertation can also be put to use in practice. Following these models, it is possible to predict the evaluation behavior of particular Internet users under particular circumstances. For instance, a high-school student who encounters a Wikipedia-article on an unfamiliar topic will most likely fail to appropriately evaluate its credibility due to a lack of domain knowledge and information skills. A postgraduate student is likely to do this better, but only because of his or her better information skills; errors on the semantic level are still likely to go undetected. The knowledge gained on how trust in information is formed can aid in the development of training and instruction in evaluation skills, as gaps in the capabilities of various Internet users can be identified better. Moreover, web developers may take user characteristics into account to develop more credible web sites, for instance by conveying positive cues about the credibility of the information presented, such as the presence of references or images.

4. Future Research

The experiments performed in this dissertation have provided substantial validation for the proposed research models. However, a few limitations should be kept in mind, which provide avenues for future studies on this topic.

In order to study the influences of various user characteristics, information characteristics, and their interplay, we introduced the Wikipedia Screening Task (Lucassen & Schraagen, 2010). This has provided valuable insights into how various factors affect the process of credibility evaluation and its outcome, trust. However, its main limitation is the isolated character of the task. First, participants performing this task were restricted to one Wikipedia article, rather than being able to search for information freely (e.g., use of different articles, search engines, or websites). Second, the sole task of the participants

was to evaluate credibility. In a real-life setting, it is much more likely that credibility evaluation is a subtask of a much larger task set. Future studies could address this limitation of the Wikipedia Screening Task, thereby verifying whether the principles captured in the studies in this dissertation are also valid in more realistic settings. However, this has implications for the extent to which the participants can be monitored, and the extent to which the stimuli can be controlled or manipulated. Moreover, it may be harder to identify whether the behavior of the participants is actually relevant for credibility evaluation, or if it is related to a different subtask (e.g., information selection).

All experiments in this dissertation were performed in the context of the free encyclopedia Wikipedia. This has provided us a great case study, as it delivered an enormous corpus of stimulus materials on virtually every topic imaginable, featuring systematically assessed information quality and high external validity for many of our participants (i.e., students; Head & Eisenberg, 2010; Lim, 2009). However, in future studies, the validity of our models should be evaluated in different (online and offline) contexts, as some findings may be specific for the context of Wikipedia. The motivation of participants to carry out their tasks systematically and extensively, for instance, may have been quite limited overall, as they were mostly asked to evaluate information of an encyclopedic nature, which mostly does not imply very high consequences of poor information. Other domains, such as the domain of medical information, may yield different results because of more severe consequences (e.g., illness or even death). Moreover, the concept of 'source' is particularly vague in the context of Wikipedia, as this can be defined as 'the authors behind the information', of which only very little is known, or 'Wikipedia in general'. Research has shown that with limited motivation, Internet users will only consider the most proximate source (Kang, Bae, Zhang, & Sundar, 2011), but it would be interesting to study credibility evaluation in environments where the source of information is easier to determine.

References

1. Publication list

Lucassen, T., & Schraagen, J. M. (2010). Trust in Wikipedia: How users trust information from an unknown source. In *Proceedings of the 4th Workshop on Information Credibility (WICOW '10) April 27 2010,* 19–26. New York, NY, USA. ACM Press.

Lucassen, T., Noordzij, M. L., & Schraagen, J. M. (2011). Reference Blindness: The influence of references on trust in Wikipedia. In *Proceedings of the ACM WebSci'11, June 14-17 2011*. Koblenz, Germany. (Chapter 4)

Lucassen, T. & Schmettow, M. (2011). Improving credibility evaluations on Wikipedia. In Wiering, C. H., Pieters, J. M., Boer, H.,editors, *Intervention Design and Evaluation in Psychology*, pages 282-308. University of Twente, Enschede, The Netherlands.

Lucassen, T. & Schraagen, J. M. (2011a). Evaluating WikiTrust: A trust support tool for Wikipedia. *First Monday, 16.*

Lucassen, T. & Schraagen, J. M. (2011b). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology, 62*, 1232-1242. (Chapter 2)

Lucassen, T. & Schraagen, J. M. (2011c). Researching Trust in Wikipedia. In *Chi Sparks, June 23, 2011*. Arnhem, The Netherlands.

Lucassen, T., Dijkstra, R.L., & Schraagen, J.M. (2012). Readability of Wikipedia. *First Monday, 17.*

Lucassen, T. & Schraagen, J. M. (2012a). Propensity to Trust and the Influence of Source and Medium Cues in Credibility Evaluation. *Journal of Information Science*, 38, p. 566-577. (Chapter 6)

Lucassen, T. & Schraagen, J. M. (2012b). The role of topic familiarity in online credibility evaluation support. In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting (HFES 2012), October 22-26, 2012*. Boston, MA, USA. (Chapter 8)

Lucassen, T., Muilwijk, R., Noordzij, M. L., & Schraagen, J. M. (2013). Topic Familiarity and Information Skills in Online Credibility Evaluation. *Journal of the American Society for Information Science and Technology.* (Chapter 3)

Van Maanen, P.-P., Lucassen, T., & van Dongen, K. (2011). Adaptive attention allocation support: Effects of system conservativeness and human competence. In Schmorrow, D. and Fidopiastis, C., editors, *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*, volume 6780 of *Lecture Notes in Computer Science*, chapter 74, pages 647-656. Springer Berlin / Heidelberg, Berlin, Heidelberg.

2. References

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 483–495.

Adler, B. T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., & Raman, V. (2008). Assigning trust to Wikipedia content (Tech. Rep. No. UCSC-SOE-08-07). School of Engineering, University of California, Santa Cruz.

Alexa Top 500 Global Sites (n.d.) In *Alexa Internet*. Retrieved September 7, 2012, from http://www.alexa.com/topsites.

Alexander, B. (2006). Web 2.0: A New Wave of Innovation for Teaching and Learning? *Educause Review, 41,* 32-44.

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

American Library Association Presidential Committee on Information Literacy. (1989). Final report. Chicago, IL.

Blumenstock, J. E. (2008). Automatically assessing the quality of Wikipedia articles (Technical Report UCB iSchool Report 2008-021). School of Information, UC Berkeley.

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.

Broadbent, D. E., Fitzgerald, P., & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the context of complex systems. *British Journal of Psychology, 77,* 33-

50.

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.

Chaiken, S. & Maheswaran, D. (1994). 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,* 460-473.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8,* 293-332.

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

Cheung, C. M. & Lee, M. K. (2001). Trust in internet shopping: Instrument development and validation through classical and modern approaches. *Journal of Global Information Management, 9,* 23-35.

Chevalier, F., Huot, S., & Fekete, J. D. (2010). WikipediaViz: Conveying article quality for casual Wikipedia readers. In *PacificVis '10: IEEE Pacific Visualization Symposium*, 215-222, IEEE.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. In *Proceedings of the 1998 Command and Control Research and Technology Symposium*, 1-37. Washington, DC, USA.

Cooke, N. J., & Breedin, S. D. (1994). Constructing naive theories of motion on-the-fly. *Memory and Cognition, 22,* 474-493.

Cormode, G. & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday, 13*.

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

Costa, P. T. & McCrea R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

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

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science, 35,* 982-1003.

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge, MA: The MIT Press.

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM, 48*, 152.

Dondio, P., Barrett, S., Weber, S., & Seigneur, J. (2006). Extracting trust from domain analysis: A case study on the Wikipedia project. In Yang, L., Jin, H., Ma, J., & Ungerer, T. (eds.), *Autonomic and Trusted Computing,* volume 4158 of *Lecture Notes in Computer Science,* chapter 35, pages 362-373. Springer Berlin / Heidelberg, Berlin, Heidelberg.

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

Dutton, W. H. & Shepherd, A. (2006). Trust in the internet as an experience technology. *Information, Communication & Society, 9,* 433-451.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, P. B. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58,* 697-718.

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*.

Ennals, R., Byler, D., Agosta, J. M., & Rosario, B. (2010). What is disputed on the web? In

Proceedings of the 4th workshop on Information credibility (WICOW '10), 67-74, New York, NY, USA. ACM Press.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: The MIT Press.

Feingold, A. (1994). Gender differences in personality: a meta-analysis. *Psychological bulletin, 116,* 429-456.

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly, 77,* 515-540.

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.

Flanagin, A. J. & Metzger, M. J. (2011). From Encyclopædia Britannica to Wikipedia. *Information, Communication & Society, 14,* 355-374.

Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. In *ACM Conference on Human Factors in Computing Systems (CHI'03), extended abstracts*, 722–723. New York, NY, USA: ACM Press.

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),* 1-15, New York, NY, USA: ACM.

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the Special Interest Group on Computer-Human Interaction (SIGCHI) at the Conference on Human Factors in Computing Systems (CHI '99)*, 80–87. New York, NY, USA: ACM Press.

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

Grabner-Kräuter, S., & Kaluscha, E. A. (2003). Empirical research in on-line trust: A review and critical assessment. *International Journal of Human–Computer Studies, 58*, 783–812.

Galton, F. (1907). The Ballot-Box. *Nature, 75,* 509-510.

Hargittai, E., Fullerton, L., Menchen-Trevino, E., & Thomas, K. Y. (2010). Trust online: Young adults' evaluation of web content. *International Journal of Communication, 4,* 468-494.

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Hayes, A. F. (2009). Beyond baron and kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76,* 408-420.

Head, A. J. & Eisenberg, M. B. (2010). How today's college students use Wikipedia for course–related research. *First Monday, 15.*

Hilburn, B., Jorna, P., Byrne, E., & Parasuraman, R. (1997): The effect of adaptive air traffic control (atc) decision aiding on controller mental workload. In Mouloua, M. & Koonce, J. M. (eds.), *Human-Automation Interaction: Research and Practice,* pages 84–91. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, England.

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

Hoffman, R. R., Lee, J. D., Woods, D. D., Shadbolt, N., Miller, J., & Bradshaw, J. M. (2009). The dynamics of trust in cyberdomains. *IEEE Intelligent Systems, 24,* 5-11.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1982). *Communication and Persuasion: Psychological Studies of Opinion Change.* Westport, Conn. : Greenwood Press Reprint, new edition.

Hovland, C. I. & Weiss, W. (1951). 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.

Johnson, T. J. & Kaye, B. K. (1998). Cruising is believing?: Comparing internet and traditional sources on media credibility measures. J*ournalism and Mass*

Communication Quarterly, 75, 325-340.

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems, 43,* 618-644.

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.

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

Kang, H., Bae, K., Zhang, S., & Sundar, S. S. (2011). Source cues in online news: Is the proximate source more powerful than distal sources? *Journalism & Mass Communication Quarterly, 88,* 719-736.

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

Kim, H. S. & Sundar, S. S. (2011). Using interface cues in online health community boards to change impressions and encourage user contribution. In *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11),* 599-608, New York, NY, USA. ACM.

Kittur, A., Suh, B., and Chi, E. H. (2008). Can you ever trust a wiki? impacting perceived trustworthiness in wikipedia. In *The 2008 ACM Conference on Computer-Supported Collaborative Work (CSCW'08),* 477-480, New York, NY, USA. ACM.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. In *Proceedings of the annual meeting of the Human Factors and Ergonomics Society,* 576–580. Santa Monica, CA. Human Factors and Ergonomics Society.

Korsgaard, T. R., & Jensen, C. D. (2009). Reengineering the Wikipedia for reputation. *Electronic Notes in Theoretical Computer Science, 244*, 81-94.

Kubiszewski, I., Noordewier, T., & Costanza, R. (2011). Perceived credibility of internet encyclopedias. *Computers & Education, 56,* 659-667

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

Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46,* 50-80.

Lee, M. K. O. & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce, 6,* 75-91.

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

Lim, S. & Kwon, N. (2010). Gender differences in information behavior concerning wikipedia, an unorthodox information source? *Library & Information Science Research, 32,* 212-220.

Liu, Z. (2004). Perceptions of credibility of scholarly information on the web. *Information Processing & Management, 40,* 1027-1038.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108,* 9020-9025.

Luyt, B., Aaron, T. C. H., Thian, L. H., & Hong, C. K. (2008). Improving Wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology, 59,* 318-330.

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*, 709-734.

McClung, H. J., Murray, R. D., & Heitlinger, L. A. (1998). The internet as a source for current patient information. *Pediatrics, 101,* e2.

Mcguinness, D. L., Zeng, H., da Silva, P. P., Ding, L., Narayanan, D., & Bhaowal, M. (2006). Investigations into trust for collaborative information repositories: A Wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web.*

McKnight, D. H., Kacmar, C. J., & Choudhury, V. (2004). Dispositional trust and distrust distinctions in predicting high- and Low-Risk internet expert advice site perceptions. *e-Service Journal, 3,* 35-58.

McKnight, D. H. & Kacmar, C. J. (2006). Factors of information credibility for an internet advice site. In *Hawaii International Conference on System Sciences (HICSS'06)*, IEEE.

Merritt, S. M. & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and History-Based trust in Human-Automation interactions. *Human Factors, 50,* 194-210.

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, 60*, 413-439.

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias. *Psychological Bulletin, 130,* 711-747.

Moinvaziri, N. (2012). WebShot [computer software]. Available from http://www.websitescreenshots.com.

Newell, A. (1982). The Knowledge Level. *Artificial Intelligence, 18,* 87-127.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Nisbett, R. E. & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. J*ournal of Personality and Social Psychology, 35,* 250-256.

Palfrey, J. & Gasser, U. *Born Digital: Understanding the First Generation of Digital Natives*. New York, NY: Basic Books.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230-253.

Preacher, K. J. & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879-891.

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, 28*.

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43,* 45-48.

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.

Schmitz, C. (2012). Limesurvey [computer software]. Available from http://www.limesurvey.org.

Self, C. C. (2009). Credibility. In M. Salwen & D. Stacks (Eds.), *An integrated approach to communication theory and research* (second edition, pp. 435-456), New York, NY: Routledge.

Sen, S. & Lerman, D. (2007). Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of Interactive Marketing, 21,* 76-94.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63,* 129-138.

Simon, H. A. (1982). *Models of Bounded Rationality, Vols. 1 and 2*. Cambridge, MA: The MIT Press.

Simon, H. A. (1997). M*odels of Bounded Rationality, Vol. 3.* Cambridge, MA: The MIT Press.

Suh, B., Chi, E. H., Kittur, A., & Pendleton, B. A. (2008). Lifting the veil: Improving accountability and social transparency in wikipedia with WikiDashboard. In *Proceedings of the 2008 conference on Human factors in computing systems (CHI'08)*, 1037-1040, New York, NY, USA: ACM Press.

Sundar, S. S. (2008) The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital Media, Youth, and Credibility* (pp. 73-100). Cambridge, MA: The MIT Press.

Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. J*ournal of the American Society of Information Science and Technology, 58,* 366-378.

Sundar, S. S. & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication, 51*, 52-72.

Statistics (n.d.). In *Wikipedia*. Retrieved March 11, 2011 from http://en.wikipedia.org/wiki/Special:Statistics.

Stvilia, B., Twidale, M. B., Gasser, L., & Smith, L. C. (2005). Information quality discussions in Wikipedia (Technical Report ISRN UIUCLIS-2005/2+CSCW). University of Illinois.

Surowiecki, J. (2005). *The Wisdom of Crowds.* Anchor.

Taraborelli, D. (2008). How the web is changing the way we trust. In *Proceedings of the 2008 conference on Current Issues in Computing and Philosophy*, 194-204. Amsterdam, The Netherlands. IOS Press.

Viégas, F., Wattenberg, M., & Kushal, D. (2004). Studying cooperation and conflict between authors with history flow visualization. In *Proceedings of the 2004 conference on Human factors in computing systems (CHI'04)*, 575-582, New York, NY, USA: ACM Press.

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.

Waters, N. L. (2007). Why you can't cite Wikipedia in my class. *Communications of the ACM, 50*,15-17.

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

White, M. P., Pahl, S., Buehner, M., & Haye, A. (2003). Trust in risky messages: The role of prior attitudes. *Risk Analysis, 23,* 717-726.

Wikipedia:Article Feedback Tool (n.d.) In *Wikipedia.* Retrieved October 5, 2012, from http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool.

Wikipedia: Version 1.0 editorial team (n.d.). In *Wikipedia*. Retrieved September 7, 2012, 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)*, 157-164, New York, NY, USA: ACM Press.

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.

Yamamoto, Y. & Tanaka, K. (2011). Enhancing credibility judgment of web search results. In *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11)*, 1235-1244, New York, NY, USA. ACM Press.

Ye, L. R. & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly, 19,* 157-172.

Nederlandstalige samenvatting

Het Internet heeft ons leven veranderd. Probeer je eens voor te stellen hoe we bepaalde klusjes deden in de tijd dat we nog niet makkelijk online konden gaan. Hoe boekten we een buitenlandse reis bijvoorbeeld? Het antwoord lijkt eenvoudig: we gingen naar het reisbureau. Maar hoe wisten we dat het aanbod wat zij deden goed was? We konden natuurlijk een paar concurrenten bezoeken, om te controleren of zij meer beenruimte of een lagere prijs konden bieden, maar het zou best wel eens kunnen zijn dat een ander reisbureau een paar kilometer verderop een nog aantrekkelijker aanbod had. In feite wisten we dat gewoon nooit zeker.

Een ander voorbeeld: Hoe kochten we ooit een gebruikte auto in de jaren tachtig of negentig? Ikzelf ben altijd erg specifiek in mijn wensen met betrekking tot merk, model en uitvoering. De tweedehands-automarkt is groot, maar hoe vonden we ooit het gewenste exemplaar zonder Internet? In feite waren we aangewezen op een paar lokale dealers en eventueel wat advertenties in de krant. Maar dit gaf je de keuze tussen enkele tientallen auto's in plaats van de duizenden mogelijkheden die we tegenwoordig hebben. Dit zijn slechts twee willekeurige voorbeelden, maar sta eens een ogenblik stil bij wat het Internet zoal veranderd heeft in ons leven, bijvoorbeeld op het gebied van werk, onderwijs, zorg, dating, winkelen, entertainment, onderzoek, enzovoorts.

Helaas brengt het Internet niet alleen voordelen met zich mee, maar ook nadelen. Zo is het Internet bijvoorbeeld niet erg goed gestructureerd. De informatie die je zoekt staat ongetwijfeld ergens, maar het zal niet altijd even eenvoudig zijn om het te vinden. De taak van informatie zoeken en selecteren vereist een volledig nieuwe set aan vaardigheden, die niet iedereen zomaar in de vingers heeft (Palfrey & Gasser, 2008). En misschien wel belangrijker: als we eenmaal de informatie gevonden hebben, hoe weten we dan dat het betrouwbaar is?

De tweede golf van Internettechnologie (ook wel bekend als Web 2.0; Cormode & Krishnamurthy, 2008) heeft deze vraag alleen maar moeilijker te beantwoorden gemaakt door Internetgebruikers de mogelijkheid te geven en aan te sporen om zelf bij te dragen aan de informatie die op Internet staat (Alexander, 2006). Dit betekent dat gevonden informatie in veel gevallen door iedereen online geplaatst kan zijn. Dus hoe weten we dan dat deze persoon goede bedoelingen heeft? Misschien wil hij of zij wel helemaal niet dat jij de juiste informatie krijgt. Denk bijvoorbeeld eens aan corrupte regimes van sommige landen, of vage handelaren op Marktplaats, die zo snel mogelijk van hun spullen af willen: de kans is groot dat hun motieven niet in ons voordeel zijn. Maar hoe identificeren we

dit soort individuen? En verder, hoe weten we dat de persoon die informatie online gezet heeft daadwerkelijk verstand heeft van het onderwerp? Vaak weten we niet eens of het een willekeurige scholier of een expert in het onderwerp van de informatie is.

Waar het op neerkomt is dat Internetgebruikers op één of andere manier een inschatting moeten maken van de betrouwbaarheid van de informatie die ze online vinden. twee potentiële problemen worden geïdentificeerd. De eerste is dat in vergelijking met het pre-Internet tijdperk, er minder professionals (zoals redacteurs, journalisten) zijn die deze taak voor je uitvoeren (Flanagin & Metzger, 2007). Dit betekent dat de verantwoordelijkheid voor het inschatten van betrouwbaarheid verplaatst is naar de eindgebruikers, die hier vaak maar weinig training in hebben gehad. Een tweede probleem is dat de meest traditionele strategie voor het inschatten van betrouwbaarheid, namelijk door te kijken naar de bron van de informatie (Chaiken & Maheswaran, 1994; Sundar, 2008), niet zo goed meer werkt. In veel gevallen is namelijk helemaal niet bekend waar informatie vandaan komt (Lim & Kwon, 2010). Bovendien zijn bronnen vaak "gelaagd", wat betekent dat informatie door meerdere bronnen gaat voordat het de eindgebruiker bereikt (Sundar & Nass, 2001; Kang, Bae, Zhang, & Sundar, 2011). Denk bijvoorbeeld aan informatie in een weblog, dat gebaseerd is op een Wikipedia-artikel, dat weer grotendeels onderbouwd wordt door een bericht op een nieuws-website. In dit soort gevallen is het lastig om te bepalen wie er verantwoordelijk is voor de betrouwbaarheid van de informatie.

De problemen die Internetgebruikers hebben met het inschatten van betrouwbaarheid van online informatie heeft geleid tot veel onderzoek in verschillende disciplines (Flanagin & Metzger, 2007). Diverse aspecten van vertrouwen in online informatie zijn onderzocht, waaronder kenmerken van de informatie (Kelton, Fleischman, & Wallace, 2008), de context (Metzger, 2007), en de gebruiker (Fogg, 2003; Sundar, 2008). Met dit proefschrift probeer ik bij te dragen aan de bestaande kennis over de invloed van karakteristieken van de gebruiker op de inschatting van betrouwbaarheid. Verschillende theorieën en modellen doen hier voorspellingen over, maar geen van alle zijn erg specifiek over hoe gebruikers met verschillende karakteristieken omgaan met de betrouwbaarheid van online informatie.

De experimenten die worden beschreven in dit proefschrift zijn allemaal uitgevoerd in de context van Wikipedia, een grote online encyclopedie die volledig door vrijwillige bijdragen van gebruikers tot stand is gekomen. Wat deze website zo interessant maakt om betrouwbaarheidsinschattingen te onderzoeken is iets wat ik regelmatig de "Wikipedia-

paradox" noem. Aan de ene kant is allang bekend dat informatie op Wikipedia over het algemeen van hoge kwaliteit is (Giles, 2005; Chesney, 2006). Aan de andere kant kun je door de open structuur van Wikipedia nooit helemaal zeker zijn van de betrouwbaarheid van de artikelen. Hierdoor zouden gebruikers van deze online encyclopedie eigenlijk altijd een inschatting moeten maken van de betrouwbaarheid van de informatie die ze vinden.

In Hoofdstuk 2 zijn we begonnen met het in kaart brengen van online betrouwbaarheidsinschattingen door drie gebruikerskarakteristieken te identificeren waarvan wij verwachtten dat ze bijzonder invloedrijk zouden zijn op de manier waarop betrouwbaarheid wordt ingeschat, namelijk domeinexpertise, informatievaardigheden, en ervaring met de bron. We hebben deze samengebracht in een nieuw model, dat we het 3S-model van vertrouwen in informatie hebben genoemd. In dit model stelden wij voor dat domeinexpertise zou leiden tot het gebruik van semantische (de 1e S) kenmerken van informatie bij inschatten van betrouwbaarheid. Daarnaast zouden informatievaardigheden leiden tot het gebruik van "oppervlakte"-kenmerken (surface features, de 2e S). Oppervlakte-kenmerken gaan over de manier waarop informatie gepresenteerd wordt, zoals bijvoorbeeld de aanwezigheid van afbeeldingen, de lengte van een tekst, of het aantal referenties. Als laatste stelden we nog dat ervaring met de bron zou leiden tot het gebruik van bron-kenmerken (source features, de 3e S). Een combinatie van de evaluatie van kenmerken in deze categorieën leidt tot een oordeel over de betrouwbaarheid van de informatie. In een online experiment lieten we experts en beginners op het gebied van autotechniek Wikipedia-artikelen zien over automotoren. De feitelijke juistheid van deze artikelen was op verschillende niveaus gemanipuleerd, met een maximum van 50% van de onderwerpen waar een fout in zat. De validiteit van het semantische deel van het 3S-model werd aangetoond doordat het vertrouwen van experts werd beïnvloed door de fouten, en het vertrouwen van beginners niet. Echter, de invloed op het vertrouwen van experts was slechts beperkt. Dit werd gezien als bewijs dat ook experts gebruik maken van oppervlakte- en bron-kenmerken. Een categorisatie van de opmerkingen die de proefpersonen maakten liet zien dat zowel experts als beginners in meer of mindere mate gebruik maakten van alle strategieën uit het 3S-model.

Het hoofddoel van Hoofdstuk 3 was om aanvullende validatie te vinden voor het 3S-model, met name voor de tweede strategie, het toepassen van informatievaardigheden. We hebben dit bereikt door studenten van verschillende niveaus (scholieren, bachelor studenten, en promovendi) Wikipedia-artikelen te laten beoordelen in een experiment waarin de proefpersonen hardop moesten denken. Er werd aangenomen dat de verschillende

groepen studenten behoorlijk zouden verschillen in hun informatievaardigheden. Wikipedia-artikelen werden per persoon uitgezocht om er voor te zorgen dat de proefpersonen veel of juist weinig over het onderwerp wisten. De resultaten waren in overeenstemming met het 3S-model. Meer bekendheid met het onderwerp (een zwakkere vorm van domeinexpertise) leidde tot meer gebruik van semantische kenmerken. Daarnaast maakten studenten met betere informatievaardigheden (bachelor studenten en promovendi) meer gebruik van oppervlakte-kenmerken dan studenten met slechtere informatievaardigheden (scholieren). In tegenstelling tot Hoofdstuk 2 werd er geen effect van bekendheid op vertrouwen gevonden, wat te verklaren is door de verschillende wijze van manipulatie in Hoofdstuk 2 en 3. In Hoofdstuk 2 werd de feitelijke correctheid gemanipuleerd, terwijl in Hoofdstuk 3 de artikelen geselecteerd waren op basis van de kwaliteitsoordelen van het Wikipedia Editorial Team. Het is te verwachten dat de kwaliteitsverschillen vooral zichtbaar zijn op het oppervlakte-niveau, aangezien de oordelen vooral een indicatie zijn van hoever het artikel gevorderd is. Feitelijke juistheid (Hoofdstuk 2) is een semantisch kenmerk. Daarom had bekendheid in Hoofdstuk 2 wel een invloed op vertrouwen, aangezien daar een semantisch kenmerk gemanipuleerd werd, en in Hoofdstuk 3 niet, omdat daar een oppervlakte-kenmerk gemanipuleerd werd. Deze uitleg wordt ook ondersteund door de observatie dat alleen het vertrouwen van studenten met betere informatievaardigheden beïnvloed werd door de kwaliteit van de informatie.

Eén specifiek oppervlakte-kenmerk bleek voor bachelor studenten bijzonder belangrijk, namelijk de referenties. Daarom hebben we in Hoofdstuk 4 het gebruik hiervan verder onderzocht. In een experiment werden de referenties van Wikipedia-artikelen op twee dimensies gemanipuleerd, namelijk kwantiteit en kwaliteit. De kwantiteit werd gemanipuleerd door het aantal referenties te variëren tussen 5 en 25. De kwaliteit werd gemanipuleerd door alle referenties te verwisselen met die van een ander, niet-gerelateerd artikel. De verwachting was dat de kwantiteitsmanipulatie meer op zou vallen dan de kwaliteitsmanipulatie, omdat er slechts een vlugge blik nodig is om dit te ontdekken (heuristische evaluatie). De kwaliteitsmanipulatie daarentegen vereist veel meer aandacht om ontdekt te worden (systematische evaluatie). De resultaten lieten zien dat beide strategieën (heuristisch en systematisch) werden toegepast door de studenten. Belangwekkend is dat de keuze tussen de strategieën afhankelijk bleek te zijn van het vertrouwen in de bron (Wikipedia). Proefpersonen met een sceptische houding ten opzichte van Wikipedia hadden de neiging om de informatie systematischer te evalueren, terwijl degenen met een minder kritische houding slechts een heuristische evaluatie uitvoerden. Bovendien bleek ondanks de grote interesse voor referenties in Hoofdstuk 3

slechts 26% van de proefpersonen de kwaliteitsmanipulatie opgemerkt te hebben. Voor de kwantiteitsmanipulatie lag dit percentage iets hoger met 39%. We noemen dit fenomeen "referentieblindheid". Referenties zijn klaarblijkelijk belangrijk voor studenten, maar zolang ze aanwezig zijn, heeft de kwaliteit (en in mindere mate de kwantiteit) maar weinig invloed.

De derde 'S' van het 3S-model, oftewel ervaring met de bron (source) was het onderwerp van studie in Hoofdstuk 5, samen met domeinexpertise. In een eerder onderzoek (Eastin, 2001) was voorgesteld dat de bron van informatie minder belangrijk is voor het inschatten van betrouwbaarheid als de gebruiker bekend is met het onderwerp. Echter, er kon geen ondersteuning worden gevonden voor deze hypothese in het experiment dat door Eastin in 2001 uitgevoerd werd. In Hoofdstuk 5 onderzoeken we deze relatie verder door proefpersonen Wikipedia-artikelen te tonen in de standaard Wikipedia-lay-out (dus met bron-kenmerken) en in een gestandaardiseerde lay-out (gebaseerd op de lay-out van Microsoft Word 2003, dus zonder bron-kenmerken). In tegenstelling tot de originele hypothese van Eastin (2001) werd het vertrouwen van proefpersonen die onbekend waren met het onderwerp niet beïnvloed door de aanwezigheid van broneigenschappen. Proefpersonen die wel bekend waren met het onderwerp werden wel beïnvloed, maar alleen wanneer de kwaliteit van de artikelen bedenkelijk was. Onder deze omstandigheden hadden zij minder vertrouwen in de informatie wanneer ze wisten dat het van Wikipedia kwam. Het verschil tussen onze resultaten en de hypothese van Eastin (2001) kan worden verklaard door het open karakter van Wikipedia. Mensen die bekend zijn met het onderwerp zouden hun eigen kennis wel eens hoger kunnen inschatten dan dat van (onbekende) anderen.

De drie gebruikerskarakteristieken uit het 3S-model werden in een breder perspectief geplaatst in Hoofdstuk 6. In dit hoofdstuk stelden we een nieuw model voor, waar in verschillende lagen de invloed van verscheidene gebruikerskarakteristieken op vertrouwen in informatie wordt uitgelegd. De kern van dit lagenmodel is vertrouwen in de informatie zelf (gebaseerd op de evaluatie van semantische en oppervlakte-kenmerken). De eerste laag hieromheen vertegenwoordigt vertrouwen in de bron van de informatie. De tweede laag is vertrouwen in het medium waarop de bron te vinden is (een generalisatie van vertrouwen in meerdere, vergelijkbare bronnen, zoals het Internet of kranten). De buitenste laag staat voor een algemene geneigdheid tot vertrouwen als karaktereigenschap. We veronderstelden dat iedere laag een directe invloed zou hebben op de aangrenzende laag. De invloed op verder weg gelegen lagen wordt gemedieerd door de tussenliggende laag of

lagen. De resultaten van een online experiment ondersteunden alle voorgestelde relaties. Daarnaast liet een post-hoc analyse zien dat gebruikers met weinig vertrouwen in de bron (in dit geval Wikipedia) geen verschil zagen tussen artikelen van hoge en lage kwaliteit, terwijl gebruikers met meer vertrouwen in de bron dit wel zagen. Dit is een indicatie voor een negativiteits-bias voor wantrouwende gebruikers, maar geen positiviteits-bias voor vertrouwende gebruikers.

Het 3S-model en andere theorieën over vertrouwen in online informatie laten zien dat gebruikers verschillende problemen hebben met het inschatten van de betrouwbaarheid. Voorbeelden zijn een tekort aan domeinexpertise of informatievaardigheden (uit het 3S-model), en motivatie of bekwaamheid (Metzger, 2007). Een mogelijke oplossing voor deze problemen is om gebruikers te ondersteunen met een beslissingsondersteunend systeem. De toepassing en acceptatie van dergelijke systemen hebben we onderzocht in Hoofdstuk 7. Drie gesimuleerde ondersteuningssystemen werden hiervoor aan proefpersonen in een experiment aangeboden. Hierbij werden twee belangrijke afwegingen onderzocht, namelijk de keuze tussen ondersteuning gebaseerd op de meningen van gebruikers of een geautomatiseerd systeem, en de keuze tussen een complex (goede prestatie, maar slecht te begrijpen) of simpel (mindere prestatie, maar beter te begrijpen) systeem. Voor wat betreft de eerste keuze konden geen eenduidige conclusies worden getrokken gebaseerd op de resultaten; beide hebben voor- en nadelen. Vertrouwen in ondersteuning op basis van eerdere gebruikers is sterk afhankelijk van het aantal gebruikers en hun geloofwaardigheid. Daarentegen is geautomatiseerde ondersteuning gevoelig voor onterecht vertrouwen, omdat de betrouwbaarheid van het advies naar alle waarschijnlijkheid nooit 100% zal zijn in complexe dynamische omgevingen zoals het Internet. De simpele geautomatiseerde ondersteuning werd al snel uitgesloten, omdat de proefpersonen de gehanteerde ondersteuningsprincipes te simplistisch vonden.

In Hoofdstuk 8 werd een directe link gelegd tussen het 3S-model en de toepassing van geautomatiseerde beslissingsondersteunende systemen. Verschillende studies (zoals Hoofdstuk 2 en 3) hebben laten zien dat domeinexperts (of gebruikers die bekend zijn met het onderwerp) de betrouwbaarheid op een andere manier inschatten dan beginners (of gebruikers die onbekend zijn met het onderwerp). Het belangrijkste verschil zit hem in het gebruik van semantische kenmerken van de informatie, wat veel meer wordt gedaan door experts. In Hoofdstuk 8 hebben we onderzocht of dit ook consequenties heeft voor de toepassing van geautomatiseerde beslissingsondersteunende systemen. Proefpersonen die al dan niet bekend waren met het onderwerp van de informatie werden

twee verschillende ondersteuningssystemen aangeboden. Het eerste systeem baseerde het advies op semantische kenmerken (typisch voor experts), terwijl het tweede systeem zich baseerde op oppervlakte-kenmerken (typisch voor beginners). Na beide systemen een paar keer uitgeprobeerd te hebben mochten de proefpersonen zelf kiezen welk systeem ze wilden gebruiken voor het laatste artikel in het experiment. Deze keuze werd gezien als een goede indicatie voor hun voorkeur. Proefpersonen die onbekend waren met het onderwerp kozen over het algemeen het systeem dat gebruik maakte van oppervlakte-kenmerken, terwijl de proefpersonen die wel bekend waren met het onderwerp geen duidelijke voorkeur hadden. De andere gehanteerde maten toonden een vergelijkbaar patroon. We concludeerden dat ontwikkelaars van dergelijke systemen zich het beste op gebruikers kunnen richten die onbekend zijn met het onderwerp. Gebruikers die wel bekend zijn met het onderwerp zijn zelf beter in staat om de betrouwbaarheid in te schatten en hebben daardoor minder behoefte aan ondersteuning.

De modellen die in dit proefschrift zijn geïntroduceerd brengen vooral meer detail ten opzichte van eerdere modellen en theorieën. Al met al kunnen we enkele conclusies trekken uit het onderzoek dat in het kader van dit proefschrift is verricht:

- Vertrouwen in informatie is aanvankelijk vooral gebaseerd op vertrouwen in de bron, wat weer gebaseerd is op vertrouwen in het medium en een algemene geneigdheid tot vertrouwen.
- Gebruikers kunnen hun vertrouwen in informatie beter afstemmen op de daadwerkelijke kwaliteit door een actieve inschatting te maken van de betrouwbaarheid.
- Karakteristieken van de gebruiker, zoals domeinexpertise, informatievaardigheden of vertrouwen in een bron, hebben een grote invloed op de manier waarop betrouwbaarheid wordt ingeschat.
- Het is essentieel om rekening te houden met gebruikerskarakteristieken in de ontwikkeling van beslissingsondersteunende systemen, bijvoorbeeld door advies te baseren op informatiekenmerken die de gebruiker zelf ook zou gebruiken.

Dankwoord

Op de voorkant van dit proefschrift staat alleen mijn eigen naam vermeld. Eigenlijk is dat onterecht, want er zijn een aantal mensen zonder wiens inzet ik nooit tot dit eindproduct gekomen zou zijn.

Met wie anders zou ik dit dankwoord beginnen dan met mijn promotor, Jan Maarten. Jouw bezielende begeleiding is doorslaggevend geweest voor het succes van dit project. Ik heb genoten van onze discussies in Enschede, Soesterberg, of over de telefoon. Deze gaven mij altijd weer de benodigde inspiratie en energie om verder te gaan. Jouw directe, maar eerlijke commentaar op minder goed werk kon erg confronterend zijn, maar je wist dat ik het kon hebben. Het lukte je vervolgens wel altijd om mij alsnog de goede kant op te sturen. Aan de andere kant was je altijd erg lovend over mij als er successen behaald werden. Dit heeft mij het zelfvertrouwen gegeven dat ik het werk aankon, en er op een juiste manier mee bezig was. Maar jouw inzet ging veel verder dan dat er van een promotor verwacht mag worden. Toen er bijvoorbeeld financiële problemen ontstonden op de afdeling waardoor reizen bemoeilijkt werd, heb jij er persoonlijk op toegezien dat een trip naar een topconferentie waar ons werk geaccepteerd was gewoon door kon gaan. Dit vond je zo belangrijk dat zelfs jouw persoonlijke budget hiervoor opgeofferd moest worden. Jan Maarten, ik wil je hartelijk bedanken voor alles wat je de afgelopen vier jaar voor mij hebt betekend en ik hoop dat mijn promotie niet het definitieve einde betekent van onze samenwerking.

Ik heb heel veel profijt gehad van de actieve bijdragen van diverse studenten aan mijn onderzoek. De studenten die specifiek hebben meegewerkt aan de diverse hoofdstukken worden aan het einde van de desbetreffende hoofdstukken al kort bedankt. Eén student leverde zelfs dermate goed werk af dat het meer dan gerechtvaardigd was om hem op te nemen als coauteur. Echter, het werk van de studenten wat het uiteindelijk niet heeft gehaald tot dit proefschrift is eigenlijk net zo belangrijk geweest. Daarom wil ik hier alle Bachelor en Master studenten die op mijn onderzoek zijn afgestudeerd bij name noemen. In alfabetische volgorde: Andreas Bremer, Kei Long Cheung, Niklas van der Golz, Tabea Hensel, Knut Jägersberg, Chris Kramer, Welmoed Looge, Ewald Maas, Rienco Muilwijk, Koen Remmerswaal, Malte Risto, en Lotta Schulze, ontzettend bedankt. Ik kon stiekem wel eens meer van jullie geleerd hebben dan jullie van mij.

Daarnaast wil ik natuurlijk alle collega's van de afdeling Cognitieve Psychologie en

Ergonomie bedanken. Jullie zorgden voor een fijne werkomgeving, waar ik vier jaar met plezier heb gezeten. Een paar mensen wil ik nog even in het bijzonder noemen. Matthijs, bedankt voor je waardevolle bijdragen aan mijn proefschrift. Ook een speciaal woord van dank voor mijn twee collega-promovendi, Marit en Jorian: Ik heb veel aan jullie gehad, omdat jullie in bijna dezelfde situatie als ik zaten. Niet alleen kwamen jullie daarom op de juiste momenten met goede ideeën, jullie zorgden ook voor de broodnodige afleiding op zijn tijd.

Beste familie en vrienden, ook voor jullie een woord van dank. Ik kon altijd bij jullie terecht met mijn eindeloze verhalen over de positieve en negatieve kanten van het onderzoek. Ik vond altijd wel een luisterend oor, en meer dan eens wisten jullie mij met een frisse blik tot nieuwe inzichten te brengen.

En dan is er natuurlijk nog mijn vrouw, Martine. Je was er altijd voor me. Samen hebben we successen gevierd en tegenslagen verwerkt. Je was minstens even begaan met mijn promotie als ikzelf. Zonder jou was me dit nooit gelukt. Ik weet dat de laatste vier jaar niet altijd even gemakkelijk waren, onder andere door het vele reizen. De perspectieven voor de toekomst zien er wat dat betreft goed uit, en ik hoop dat ik je nu meer de aandacht kan geven die je verdient.

Curriculum Vitae

Teun Lucassen werd op 12 april 1983 in Meppel geboren. In 2001 behaalde hij zijn VWO-diploma aan de Christelijke Scholengemeenschap Dingstede in Meppel. In datzelfde jaar startte hij met de bachelor opleiding Technische Informatica aan de Universiteit Twente, welke hij in 2006 met goed gevolg afrondde. Hierna begon hij aan de master Human Media Interaction aan dezelfde universiteit. Deze studie werd in 2009 afgerond met een externe stage en afstudeeropdracht bij TNO in Soesterberg. Direct hierop volgend begon Teun aan zijn promotieonderzoek onder leiding van prof. dr. Jan Maarten Schraagen bij de vakgroep Cognitieve Psychologie en Ergonomie van de faculteit Gedragswetenschappen aan de Universiteit Twente. Tijdens zijn promotieonderzoek heeft hij werk gepresenteerd op internationale congressen, en artikelen gepubliceerd in internationaal hoog aangeschreven tijdschriften. Daarnaast is er diverse keren op nationaal niveau media-aandacht geweest voor zijn onderzoek. Teun zal zijn carrière voortzetten als Hogeschooldocent Serious Gaming aan de Hogeschool Windesheim in Zwolle.