

Trust in online information

A comparison among high school students, college students and PhD students with regard to trust in Wikipedia

Masterthesis

Rienco Muilwijk

First supervisor: Teun Lucassen, MSc

Second supervisor: Prof. dr. Jan Maarten Schraagen

Department of Cognitive Psychology and Ergonomics,
University of Twente

Trust in online information

*A comparison among high school students, college students and PhD students
with regard to trust in Wikipedia*

Rienco Mulwijk

University of Twente, February 2012

ABSTRACT

With the advent of the World Wide Web, it has become easy to obtain more information in less time. Accessibility, quantity and speed have been improved in the past years, but what about information quality? The current study focuses on how users perceive the trustworthiness of information. Three user groups, namely high school students, college students, and PhD students, made trust judgments on Wikipedia-articles which varied in quality and familiarity to the user. These three user groups were selected because as a consequence of differences in age and education progression, they are expected to differ as well in the development of information problem solving skills.

User's information problem solving skills determine, together with source experience and domain expertise, a trust judgment. These three user characteristics can be found in the 3S-model (Lucassen & Schraagen, 2011), which differentiates three strategies (source, surface, and semantics) applied by a user to evaluate information's credibility. Two relations of the 3S-model were tested in this study: on the one hand the relation between the application of a surface strategy and the degree of information skills (knowledge how to evaluate online information, Metzger, 2007); on the other hand, the relation between the application semantic strategy and the degree of domain expertise. Through the think aloud method participants indicated which information features (e.g., authority, accuracy, completeness, length) they attended to while making trust judgments. Based on the 3S-model, a coding scheme was developed. For each group, all remarks were coded and counted in order to include them in the coding scheme. The coded remarks were compared to each other to find differences and similarities between the user groups in feature and strategy application.

Results show that high school students differed from college students and PhD students in feature and strategy application. High school students frequently mentioned the accuracy of information whereby semantic strategy was applied. College students and PhD students predominantly attended to authority. Presumably, their information skills enabled them to apply a surface strategy more than a semantic strategy. As expected, all three groups have in common that they apply more semantic strategy when confronted with familiar topics compared with unfamiliar topics.

1. INTRODUCTION

Relevance

The World Wide Web provides a lot of opportunities to satisfy one's demand for communication and entertainment. Another purpose that the World Wide Web fulfills is provision of online information. Information can be found anywhere on the internet: news items are published on broadcast and newspaper websites, product information is consulted on e-commerce websites, and factual knowledge is spread via encyclopedic websites. Much information is available, but when users are interested in high quality information, not all information is usable, "since there is no quality control on how webpages get created and maintained" (Zhu & Gaugh, 2000). How do users find out whether information is of high or low quality? The current study compares three user groups who have to cope with this issue almost daily. These groups are high school students, college students, and PhD students. The main difference between the groups is their education progress. It is assumed that education benefits the development of information problem solving skills (information skills). After all, information skills, defined as "the knowledge how to evaluate online information" (Metzger, 2007) "can be characterized as a complex cognitive skill" and need "explicit and intensive instruction" to acquire (Brand-Gruwel, Wopereis, & Vermetten, 2005). As education progress increases from high school student to PhD student, the three user groups are expected to differ in their development of information skills. Differences in information skills probably result in different trust judgments. Finding out how trustworthiness of online information is perceived, or more specifically, what the effect of information skills is on trust judgment, provides valuable information for different fields. Teachers can learn more about how their students perceive information and eventually guide them more effectively in consideration which information to use and which not. In a broader perspective, this study contributes to the research on 'information-seeking behavior', exploring how humans search, acquire, process, organize and present information. For pedagogy and developmental psychology, the comparison of three different age groups can be useful for research on capabilities in different ages. In marketing and media psychology, the findings of this study may be of interest to choose the right presentation for a certain target group to come across as trustworthy. That is, research on information behavior concerns both intentional and unintentional behavior (Wilson, 1999).

What makes this study unique is that credibility evaluations of three (instead of two) user groups are compared and that these groups are selected on presumed differences in information skills. To provide insight into differences and similarities in their ways of perceiving trustworthiness, a coding scheme tailored to information quality will be designed and used.

Through related research where theoretical concepts such as information quality and trust are explained, and where relevant models are reviewed, predictions are made how different user groups evaluate the

credibility of information. Then the experiment of this study is presented, followed by its results. In conclusion, the results are reflected and interpreted.

Clarification of concepts

Because in the literature the concepts trust judgment and related concepts are defined differently by different authors, these terms are clarified to get a clear view of what is meant by those terms. Below is described how trust judgment, information quality, credibility, trustworthiness and expertise are defined and relate to each other.

To start with, there is trust judgment. Chopra and Wallace (2003) collected from literature frequently used definitions of trust. They summarized them into one definition of trust: “Trust is the willingness to rely on a specific other, based on confidence that one’s trust will lead to positive outcomes.” Chopra and Wallace also mentioned dependency and risk as two preconditions for trust. In their definition, trust is considered an individual’s attitude. This attitude is directed toward someone or something (e.g. information). Depending on the quality of information, trust mediates if information is used or not (Kelton, Fleischmann, & Wallace, 2008). Trusting and using information means one has to take a risk (Johnson-George & Swap, 1982) because it is uncertain if the expectation of positive outcome and confidence in the judgment (Blomqvist, 1997; Giddens, 1990) are justified. However, the usage of information does not only depend on perceived information quality. Another determiner to use information is perceived credibility. Information quality and credibility are discussed below.

Information quality (IQ) is the actual quality of an information piece. Information consists of many different features that all together determine its quality. To make a trust judgment on information, it is helpful to know which features are most relevant to take into account. Different researchers tried to frame the most important information features. For example, Alexander and Tate (1999) mentioned objectivity, completeness and pluralism. Others included accuracy, authority, and currency. In total, 27 different features were found in 18 studies. Table 1 lists the top 10 of the most mentioned features. As can be concluded from the list, especially currency, accuracy, and authority are important determiners of information quality and thus essential features to take into account when making a trust judgment.

Credibility is the believability of information (Hovland, Janis, & Kelley, 1953), consisting of two components: trustworthiness and expertise (e.g. Fogg & Tseng, 1999; Metzger, 2007). Fogg and Tseng (1999) define trustworthiness as “the perceived morality of a source.” It deals with the question to which extent a source is for example well-intentioned, truthful, and unbiased. Expertise as a component of credibility is “the perceived knowledge and skill of the source”. Besides perceived trustworthiness and perceived expertise, source attractiveness and dynamism can also influence the perception of credibility (O’Keefe, 2002). Since evaluation of quality and credibility depend on a lot of factors, and users can be

Table 1

Approach of information quality

Feature ^a	Description	N Studies ^b
Currency	Content up-to-date	14
Accuracy	Correctness, absence of errors	13
Authority	Support of relevant literature	11
Completeness	Presence of major facts and details and contextual placement of the subject	7
Objectivity	Content is written from a neutral point of view	7
Stability	Frequency content edits	7
Structure	Presentation of the content (e.g. presence of headings)	6
Writing style	Professional and readable language	6
Length	Number of words	5
Availability	Number of broken links by total number of links	5

Note: Descriptions and counts are based on: Alexander & Tate (1999), Ballou, Wang, & Kumar (1998), Cassel (1995), Chopra & Wallace (2003), Crawford (2001), Dondio, Barret, Weber, & Seigneur (2006), Emigh & Herring (2005), Eppler & Muenzenmayer (2002), Kelton et al. (2008), Lih (2004), Lim (2009), Motro & Rakov (1998), Naumann & Rolker (2000), Stanford Web Credibility (2005), Stvilia, Twidale, Smith, & Gasser (2005), Viégas, Wattenberg, & Dave (2004), Wang, Allen, Harris, & Madnick (2002), Zhu & Gaugh (2000).

^a Only the most mentioned features (5 studies or more) are presented here (10 out of 27).

^b Total number of studies is 18.

biased in their approach (Fogg & Tseng, 1999), persons make different trust judgments toward the same piece of information. Therefore, when a user evaluates IQ or credibility, his review is always *perceived* IQ and *perceived* credibility.

In this study, the terms information quality, information credibility, and trust judgment are used. Hereby, information quality refers to an approach to the actual quality of Wikipedia articles (approached by Wikipedia itself, later explained in more detail), information credibility is a theoretical term used to describe the actual believability of information, and trust judgment is defined as a participants' evaluation of the trustworthiness of information, including the consideration whether to use the information or not.

Now that the concepts are clarified, research on trust judgment is discussed next.

Related work

Lucassen and Schraagen (2011) developed a model with three strategies being applied when a user makes a trust judgment. This '3S-model' (Figure 1) distinguishes between semantic features, which concern content of information; surface features, regarding how information is presented; and source features, about the reputation of the source where the information is found. The 3S-model is discussed here, for

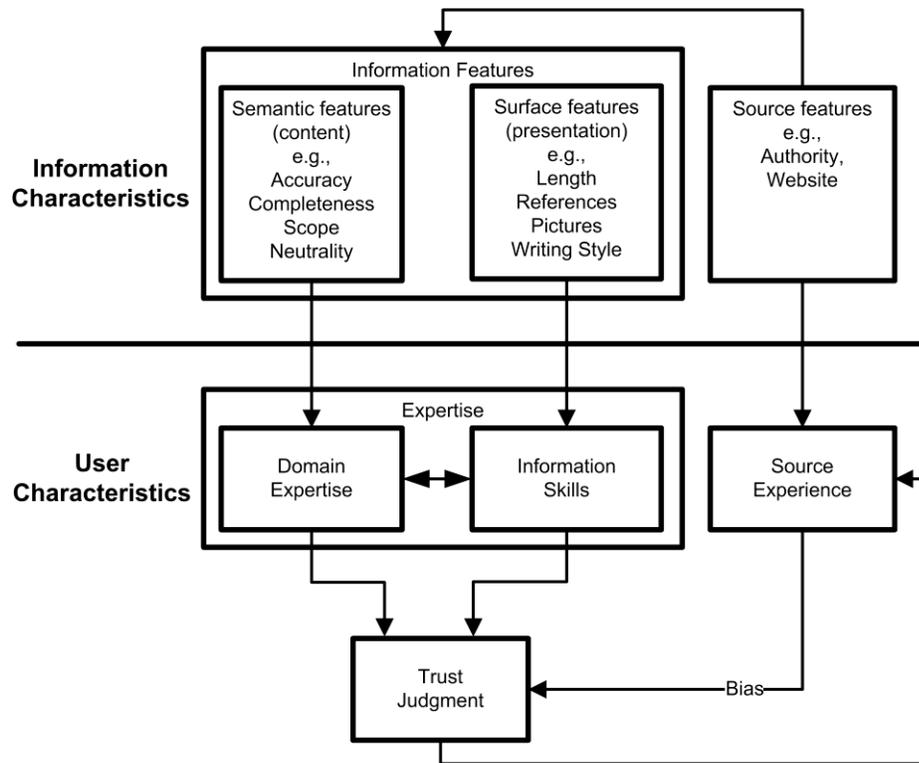


Figure 1. The 3S-model of information trust (Lucassen & Schraagen, 2011).

differences in trust judgment between the three user groups of the current study might be explained by differences in the extent to which strategies are applied. The three strategies are described in more detail below.

Users attend to semantic features for checking information on its content. Examples of semantic features are accuracy, completeness, and objectivity (neutrality). According to Lucassen and Schraagen, domain experts (i.e. people with thorough knowledge about a certain topic) are more able to consider content in their trust judgment than novices, because where domain experts can compare presented information that is familiar to them with their knowledge, novices cannot make such a comparison.

The presentation of information (surface) is another strategy to approach information credibility (Chevalier, Huot, & Fekete, 2010; Fogg & Tseng, 1999; Fogg, Danielson, Marable, Stanford, & Tauber, 2003). In fact, surface strategy is the first impression of information credibility (Walthen & Burkell, 2002). Online information can be presented in very different ways, like writing style, presence of images, and length of a text. Some users possess more information skills than others, presumably because of education level and progress (Brand-Gruwel et al., 2005).

When users already have experience with a source of information, this also influences the outcome of their trust judgment. “I do not trust anything from Wikipedia” is a quote that fits in a source strategy. Someone’s perception about the quality of a source is called ‘online reputation’ by Simpson (2010).

The 3S-model is based on the Dual Processing Model of Website Credibility (Metzger, 2007). Metzger states that the approach toward trust depends on user's motivation and ability. Motivation to investigate information credibility decreases when a user already has experience (positive or negative) with a source (Lucassen & Schraagen, 2011). Ability is expressed in the 3S-Model by 'user characteristics': a trust judgment is based on user's knowledge of a specified topic (domain expertise), his knowledge about considerable parts of information (information skills), and his experience with a certain source (source experience).

The model described above resembles the 'Model for Ascribing Cognitive Authority to Internet Information', proposed by Fritch and Cromwell (2001). Fritch and Cromwell focused on cognitive authority, which is defined as "influence on one's thoughts that one would consciously recognize as proper" (Wilson, 1983). In their model, four different classes of cognitive authority are presented: document, author, institution, and affiliation. Firstly, document authority is based on both presentation and content of information and largely overlaps with the semantic and surface strategies from the 3S-model. Secondly, author authority is determined by who has written the information. Thirdly, institution authority is the reputation of the source presenting the information. Lastly, affiliation authority describes the presence of affiliations between source of information and other parties. The last three classes correspond with source strategy from the 3S-model.

Besides a distinction in strategies upon which a trust judgment is based, different psychosocial levels of trust exist. Kelton et al. (2008) distinguish four levels of trust: individual (a personal characteristic), interpersonal (a one way direction from one person to another), relational (a mutual direction between two or more persons), and societal (trust of one or more persons in a community). One's trust judgment in information is a form of interpersonal trust, because trusting information implies trusting its author (Lucassen & Schraagen, 2011). In the situation of online information it is not always clear who the author is. An example is Wikipedia.

Wikipedia

For the reasons that Wikipedia's information quality is high, its quantity is large, its range is wide, this online encyclopedia has been found most suitable for studying trust judgment. After a short introduction, more is explained about Wikipedia's information in terms of quality, quantity, and diversity.

Launched in January 2001, the free encyclopedia Wikipedia has risen in popularity, ranking 6th of the most popular websites in the world¹ and is found in top positions by results of search engines. Wikipedia not only provides encyclopedic facts, it also offers anyone the opportunity to change and add articles.

¹ www.alexa.com/topsites

This ‘open character’ makes it possible for users to actualize articles continuously and in addition, the knowledge comes from all over the world. Unfortunately, the drawback is Wikipedia’s vulnerability to manipulation and vandalism. User’s contributions are seen both as strength and threat to the content (Viégas et al., 2004).

The quality of articles is of fundamental importance to any encyclopedia. Although Giles (2005) showed that Wikipedia’s IQ is comparable with Britannica Encyclopedia’s, in that same year former USA Today journalist John Seigenthaler found himself a victim of a hoax on his Wikipedia biography page. In his criticism about Wikipedia, Seigenthaler predicted a massive growth of vandalism which would result in the need for government regulation of the encyclopedia (Seigenthaler, 2005).

The International Association for Information and Data Quality (IAIDQ) distinguishes three states in the process of making information high qualitative: assessment, control and improvement (English, 2005). As the users perform the part of information assessment, Wikipedia finds itself active in checking articles for incorrectness and in improving from a ‘stub’ status, that is when an article is only a basic description of the topic, ultimately to a ‘featured article’ (FA), which is an article containing very professional information. Wikipedia’s level of quality depends on the arrangement of (1) determining the current quality level, (2) improving the quality level, and (3) preserving the high quality. The assignment of a rating is done by the ‘Wikipedia: Version 1.0 Editorial Team’² and is called the WP 1.0 program. Rating is also performed by the so-called project members (a group of Wikipedia editors). IQ of a Wikipedia article is determined by: accuracy of information, appropriateness of the images, appropriateness of the style and focus, susceptibility to false information, comprehensiveness, identification of reputable third-party sources as citations, stability, susceptibility to editorial and systemic bias, and quality of writing³. These criteria correspond with the 10 earlier mentioned relevant information features from the literature (see Table 1) and add images to it. The literature and Wikipedia together bring forward 11 relevant information features to evaluate information quality. Relevant information features mentioned by literature refer to those 11 features.

Given that the English Wikipedia has over 3.8 million articles⁴, its large quantity provides a high diversity of topics. Thanks to this diversity, every user can find articles which are familiar and unfamiliar to him. This turned out very useful in the selection of articles for the participants in this study (see method section).

² http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team

³ http://en.wikipedia.org/wiki/Reliability_of_Wikipedia

⁴ <http://stats.wikimedia.org/EN/TablesWikipediaEN.htm>

Hypotheses

In reference to the previous section, researchers and the Wikipedia Editorial Team specified the most relevant features of information quality. Now that literature's perception on IQ features is reviewed, what follows is a closer look at user's perception of IQ. How IQ is perceived by users has been studied in general (no specific user group) by Fogg et al. (2003), for Wikipedia members by Yang and Lai (2010), for high school students by Lorenzen (2002), and Walraven, Brand-Gruwel, and Boshuizen (2009), for college students by Lucassen and Schraagen (2010), Menchen-Trevino, and Hargittai, (2011), and for both college students and PhD students by Brand-Gruwel et al. (2005). The current study is the first to compare not just two groups, but three: high school students, college students, and PhD students. The applied information features of each group (white circles in Figure 2) will be compared with the information features selected by the literature and Wikipedia (black circles in Figure 2) to find out to which extent different user groups agree with the information features from the literature. As stated earlier, the three user groups differ in education progress and therefore it is assumed that they also differ in the possession of information skills. Grimes & Boening (2001) found for example that high school students selected unauthorized resources. Brand-Gruwel et al. (2005) selected first-year college students as information novices and PhD students as information experts. Differences in the application of

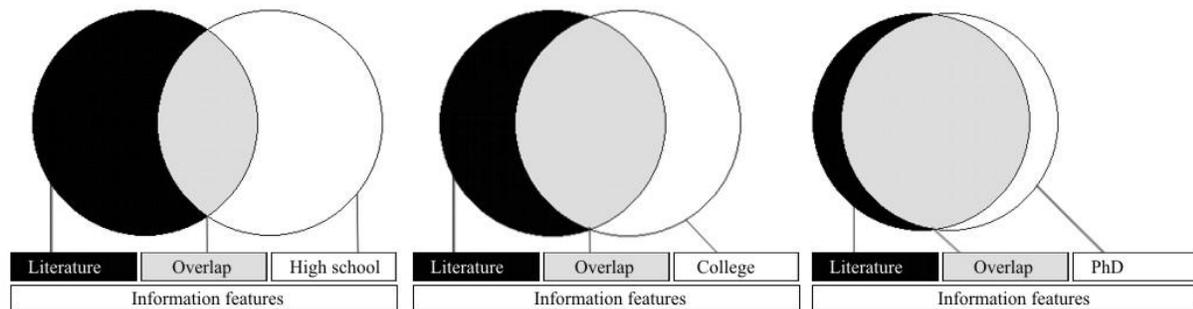


Figure 2. Expected distribution of applied information features by high school students, college students, and PhD students.

information features are expected between groups. This expectation, which is the first hypothesis, is visualized in Figure 2. The figure displays the expected distribution of how the user groups will apply information features with respect to the 11 relevant information features mentioned in the literature (Table 1). The gray area is the overlapping area, which means that these features are mentioned both in the literature and by the user group. As can be seen, it is expected that the size of this area increases as the user group has more information skills.

Hypothesis 1: PhD students apply more relevant information features than college students and high school students when they judge trustworthiness of a Wikipedia article; college students apply more relevant information features than high school students.

Compared to the other two groups, PhD students are likely to have the most experience in processing information. PhD students are considered information specialists or at least they are the closest to that. To become an information specialist, one has to acquire certain skills; in that case, the user has a sufficient set of tools. When these skills are not acquired, one has to make a trust judgment with a toolbox that is incomplete. The 3S-model links information skills to surface strategy. Consequently, it is expected that the application of surface strategy increases as more information skills are acquired. Hence, it is predicted by Lucassen and Schraagen (2011) that information specialists will largely apply their information skills. This prediction was based on Brand-Gruwel et al. (2005), who found differences in trust judging behavior between (information) experts and novices. This expectation, together with the assumption that there is an upward trend in information skills from high school students to PhD students, leads to the following hypothesis:

Hypothesis 2: PhD students will base their trust judgment on surface features more than college students and high school students. College students will base their trust judgment on surface features more than high school students

The third hypothesis continues with predicting how strategies probably will be applied. When the 3S-model was explained, it was already mentioned that domain expertise enables the use of semantic features. That is, when users know much about a certain topic, they are likely to base their trust judgment among other things on accuracy and completeness. This idea was already tested before the 3S-model was introduced: Lucassen and Schraagen (2011) referred to the study of Chi, Feltovich, and Glaser (1981) who found that novices (i.e. people without or only little knowledge about a topic) applied surface features to sort physics problems while (domain) experts categorized the problems, which can be interpreted as a semantic way to overcome a problem. Another study (Adelson, 1984) indicated that knowledge about something is positively related to the application of semantic information features. She showed that in the domain of computer programming, novices were better in answering concrete questions (surface), whereas experts answered abstract question better (semantics). Translating these findings to trust judgment of online information, the more knowledge users have about a certain topic (hence, domain familiarity), the more they are likely to pay attention to semantic features. To illustrate this: a piece of information with many inaccuracies and factual errors is likely to be rejected by users who are familiar with that topic. Thinking that something is incorrect means one has at least an idea what a topic should be about and that the presented information does not correspond with this idea. All together, hypothesis 3 reads as follows:

Hypothesis 3: Semantic features are applied more on familiar topics than on unfamiliar topics.

2. METHOD

Participants

On the basis of education progression, participants ($N = 40$) can be distinguished in three groups. The first group consisted of high school students ($n = 13$; 5 male and 8 female) who were all receiving pre-academic education. The mean age for high school students was 14.3 years ($SD = 0.6$). Group 2 were college students ($n = 12$; 5 male and 7 female). In this group, seven were Dutch and 5 German. Their mean age was 23.4 years ($SD = 6.3$). The third group consisted of PhD students ($n = 15$; 7 male and 8 female) with a mean age of 27.0 years ($SD = 1.9$). Their research disciplines differed so that both theoretical as well as applied sciences were represented.

All participants were familiar with Wikipedia, using the encyclopedia for 2-10 years. College students and PhD students understood the principle of Wikipedia as an open source which anyone can edit; however, when high school students were asked what characterizes Wikipedia, most of their answers did not cover what Wikipedia stands for.

As a sequel to Lucassen and Schraagen (2010), transcriptions from the college students were already available.

Task

The Wikipedia Screening Task (WST, Lucassen & Schraagen, 2010) was performed. Preceding the start of the task, the experimenter instructed the participant to focus on the trustworthiness of each article and to think aloud while performing the WST. During the task, the experimenter was silent.

To participants in the experiment subsequently 10 Wikipedia-articles were presented. These articles were presented offline, because small manipulations to the articles had been applied. That is, cues of trustworthiness or quality were removed (for example the call ‘citation needed’ and a star in the upper right corner indicating that an article has ‘feature article status’).

The absence of a time limit and the correspondence of article’s language to the reading knowledge of the participant (Dutch to the high school students and English to the university and PhD students) made the task manageable for all participants. All participants were restricted to the displayed article, which meant that they were not permitted to follow internal links nor visit any other website.

Design

With three independent variables, a 3 (student group: high school students, college students, and PhD students) x 2 (familiarity: familiar and unfamiliar) x 2 (article quality: high and low) design was set up. Three different groups of students, differentiated by education progress, participated who read both

articles with a familiar and unfamiliar topic of articles of which some of high quality and others low quality. The variables familiarity and quality of articles are discussed below.

To find out participants' familiarity to certain topics, they were briefly interviewed. Based on these interviews five familiar and five unfamiliar topics were chosen for each participant. Wikipedia-articles of those topics were found, downloaded and manipulated with regard to cues about quality and credibility. To check if the assumed familiarity matched with the real familiarity of a participant with an article, participants rated familiarity after each article on a 7-point Likert scale, ranging from totally unfamiliar to totally familiar with the topic.

When a Wikipedia-article has a quality rating, it can be found on Wikipedia's discussion page. These ratings, invisible during the WST, were used to select both low quality and high quality articles. The Wikipedia Editorial Team's ratings range from 'stub' when an article is only a basic description of the topic to 'featured article' (FA) for very professional information. Further specification per class (or status) is included in Table 2, and more details (for example reader's experience and editing suggestions) can be found on a Wikipedia page devoted to the different classes⁵. Since there is an underrepresentation of A-

Table 2

Wikipedia Editorial Team assessment

Class	Criteria
FA	Attainment of featured article status.
A	Well-organized and essentially complete, having been reviewed by impartial reviewers from a WikiProject. Good article status is not a requirement for A-Class.
GA	Attainment of good article status.
B	Mostly complete, without major issues, some further work to reach good article standards is required.
C	Substantial, but is still missing important content or contains a lot of irrelevant material. The article should have references to reliable sources, but may still have significant issues or require substantial cleanup.
Start	In development, quite incomplete, further reliable sources required.
Stub	A very basic description of the topic.

⁵en.wikipedia.org/wiki/Wikipedia:1.0/A

class articles, those articles consequently are not as representative as articles in other categories. Besides that, an underrepresentation of A-class articles restricts the experimenter in choosing familiar and unfamiliar topics. For those two arguments, A-class articles were excluded. Six categories remained, of which the three highest classes were considered high quality (FA, GA, and B) and the three lowest classes were considered low quality (C, Start, and Stub). Half of the presented articles (five) were high quality; the other half was low quality. Article quality was randomized between trials. Participants made a trust judgment to each article (thus ten trust judgments per participant) on a 7-point Likert scale ranging from totally not trustworthy to totally trustworthy information.

Procedure

One participant per session took part in the experiment. After attending to the informed consent, and after filling in questionnaires on demographics and familiarity with Wikipedia, instructions were given how to think aloud. Then the participants practiced the task by evaluating two practice articles. Making a judgment on an article's credibility had no time limit and was noted by each participant on a questionnaire. During these practice trials, the experimenter provided feedback. All participants passed the practice trials properly. Then ten articles were presented in succession, but the experimenter did not comment on participant's performance anymore. Only when a participant fell silent, the experimenter stimulated him to keep thinking aloud. Participants took roughly 90 minutes to complete a session.

Analysis

Every session was sound-recorded, which made it possible to transcribe, select, and code the selected phrases. Phrases were selected when the question: "Does this comment involve credibility evaluation?" could be answered positively. Phrases beyond the scope of credibility were not taken into account for analysis.

To code the selected phrases, a coding scheme was developed. The foundation of the coding scheme, shown in Figure 3, is the 3S-Model. In this model, Lucassen and Schraagen (2011) distinguish three strategies which do not exclude each other's application: a user may apply more than one strategy to evaluate. The experiment's context is that of Wikipedia only, therefore no source comments were expected. This is why 'source' does not appear in the coding scheme. The remaining two strategies, surface and semantics, were included in the coding scheme, recurring at every component of an article.

	Article											
Component	Introduction		Text		Table of contents		Images		References		Other	
Strategy	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics
Feature												

Figure 3. Coding scheme based on the 3S-Model (Lucassen & Schraagen, 2011). In this scheme, 'images' is considered a component instead of a feature. In the analysis, all features of images are taken together. The same goes for 'references' (authority). The 'source' strategy is not considered as no comments about Wikipedia are expected within the context of the WST.

In the coding process, first the selected phrases (remarks) were categorized by component (e.g. introduction, images). Then it was specified which feature (e.g. number, quality) was meant. Thirdly, the applied strategy was coded (surface or semantics). The coded remarks were counted, but before they could be analyzed, they were corrected for verbal fluency (total number of remarks made). To prevent disproportional influence of some participants compared to others, a correction factor was calculated for each participant. The corrected number for a participant was computed by dividing the total number of remarks within a group by the number of participants in that group and subsequently this fraction was divided by the number of remarks for one participant.

$$\text{corrected number} = \frac{\text{total number of remarks / number of participants}}{\text{number of remarks}} \quad (1)$$

To ensure the reliability of the coding of the protocols, the inter-rater reliability between both experimenters for all three user groups was calculated by double-coding one protocol from the other experimenter. Results of the inter-rater reliability analysis show near perfect agreement between the experimenters ($\kappa = .87$).

3. RESULTS

Manipulation check on familiarity

Half of the Wikipedia articles were assumed to be familiar to the participants, whereas the other half concerned unfamiliar topics. To check if the participants indeed were familiar or unfamiliar with a selected topic, they were asked to rate each article on familiarity subsequently to the trust judgment. The assumed familiarity was compared with this familiarity score and the assumption turned out to be valid for all three groups: high school students ($z = 3.78, p < .01$); college students ($z = 4.17, p < .01$); and PhD students ($z = 4.68, p < .01$).

Applied information features: user groups and the literature compared

Hypothesis 1 was tested to compare each of the three user groups' applied information features with the 11 relevant information features mentioned in the literature and described in Table 1. The participants mentioned 23 different information features. Some of these information features correspond to the relevant information properties (from literature). Participants also mentioned other information features, the 'supplementary features', which are not prominently present in the literature. These supplementary information features are explained and then the results of hypothesis 1 are discussed.

Participants mentioned 'quality', which refers to the appearance of a component. "This text looks nice," is an example of an information feature that was coded for quality. The feature 'scope' is about the range of a component, about whether the information is to the point. Participants also made comments about the number (amount) of for example images and the presence of for example references. 'Factual accuracy' in the coding scheme and 'accuracy' in the literature are the same, as 'references' and 'authority' are the same.

To compare the users' application of information features with the relevant information features according to the literature, a classification of three partitions was distinguished. For each group, an applied information feature was classified as *overlapping* when it corresponded with a relevant information feature according to the literature. An information feature was classified as *supplementary* when that feature was not mentioned in the literature. The third partition consists of information features that were mentioned in the *literature*, but not by the user group. The overlapping and supplementary information features had to be mentioned by a majority of participants within a user group to be taken into account for the comparison.

In reference to Figure 2, where a distribution was predicted, Figure 4 displays the distribution how it was found. The proportion of overlapping features (gray area of the figure) was compared with the proportion of relevant features according to the literature (black area of the figure) and the proportion of features that were supplemented by the participants (white area of the figure).

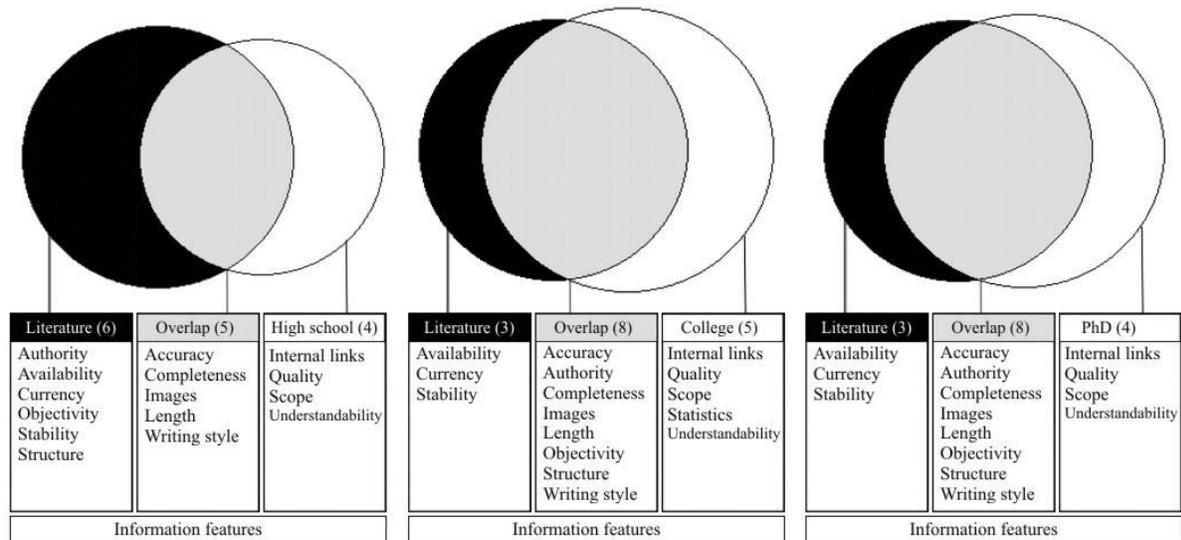


Figure 4. Distribution of applied information features by high school students, college students, and PhD students. High school students mentioned 9 different features, of which 5 are seen as most relevant according to literature. This proportion was 8/13 for college students and 8/12 for PhD students. Due to differences in the number of unique features mentioned per group, the circles differ in size. Features were included in the diagram when they were mentioned by a majority of participants within a group.

High school students made remarks about accuracy of text, completeness of text, images, length of text, and writing style, which are five information features that are also mentioned in the literature as relevant to judge information quality. Besides these five overlapping information features, high school students applied four supplementary information features. As can be seen in the figure, the overlapping area of information features between the literature and the user group is smaller for high school students (five) than for college students and PhD students (both eight). The last two groups attended to the same relevant properties as the high school students and in addition they applied authority, objectivity, and structure.

Another difference between high school students on the one hand and college students and PhD students on the other hand is that the last two groups applied more different information features. In Figure 4, this difference is illustrated by a larger white circle and a longer list in the right box, containing supplementary information features. High school students applied nine different information features (five overlapping features and four supplementary features); college students applied 13 and PhD students 12 information features. The group of college students is the only one that takes statistics into account.

Hypothesis 1 is not fully supported. On the one hand there is a difference between high school students and the other two groups. On the other hand, no significant differences were found between the application of information features by PhD students and college students.

From the coding schemes, it turned out that the number of remarks on the feature 'accuracy of text' dropped considerably when education progression increased. High school students made 38.2 % of their remarks about the accuracy of text; college students did this for only 12.7 % of their remarks and PhD

students mentioned accuracy 12.6 % of their total remarks. An opposite relation goes for authority: it had very few remarks from high school students (0.5 %), but it was by far the most prominent feature for college students (28.4 %) and PhD students (33.4 %). The application of accuracy and authority per group are presented in Figure 5. Because participants were not permitted to follow links, they could not check availability. Neither could they check stability, because this feature is not visible at the ‘read page’ of a Wikipedia article.

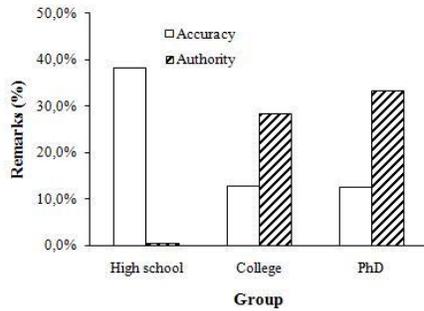


Figure 5. The most applied information feature for high school students (accuracy) against the most applied information feature for college students and PhD students (authority).

Strategy application between groups

The second hypothesis was tested to find differences and similarities in strategy application between the three user groups. Two out of three strategies from the 3S-model (Lucassen & Schraagen, 2011) were expected: surface and semantics, whereby it was hypothesized that PhD students would apply more surface strategy than the other two groups.

An overview of strategy application per group is given in Table 3 and Figure 6. The groups differed from each other in strategy application: $\chi^2(2, N = 2793) = 111.68; p < .01$. College students made surface remarks the most (67.2 %), followed by the PhD's (63.7 %), and high school students (42.8 %). The difference turned out to be significant only if a group was compared with the high school students. Thus, PhD's, $\chi^2(1, N = 1671) = 70.98; p < .01$ and college students, $\chi^2(1, N = 1783) = 101.66; p < .01$ made more surface based remarks than high school students. These results support hypothesis 2.

Table 3

Strategy application

Strategy	High school	College	PhD
Surface	42.8 %	67.2 %	63.7 %
Semantics	57.2 %	32.8 %	36.3 %

Note: Percentages are based on corrected number of remarks.

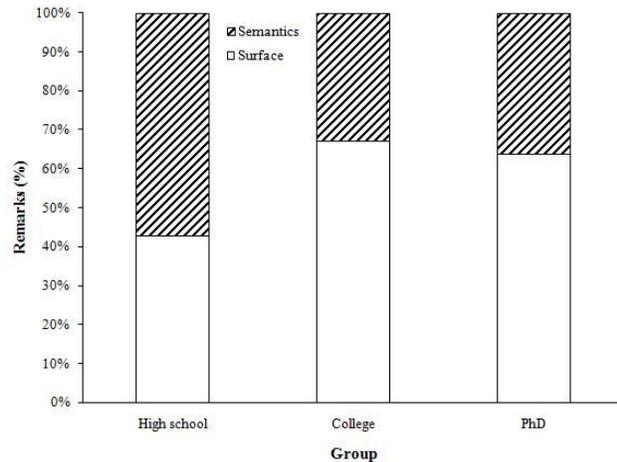


Figure 6. Percentages of strategy application for high school students, college students, and PhD students.

Strategy application and familiarity

Hypothesis 3 focuses on which strategy is applied most when an article is familiar to the user. It was expected that participants would apply the semantic strategy more when they were familiar with a topic. To test this hypothesis, all semantic remarks were counted. This was done with respect to two categories: one sum for all familiar topics and one sum for all unfamiliar topics. The distribution is shown in Table 4.

Table 4

Strategy application

Strategy	Familiar	Unfamiliar
Surface	52.9 %	68.7 %
Semantics	47.1 %	31.3 %

Strategy application of high school students, college students and PhD students taken together.

Note: Percentages are based on corrected number of remarks.

Table 4 shows that when users are familiar to a topic they apply a semantic strategy in 47.1 % of their total remarks; this is more than the 31.3 % of semantic remarks on unfamiliar topics. The differences in strategy application between familiar and unfamiliar topics are significant with $\chi^2(1, N = 2793) = 72.38$; $p < .01$). Indeed, results support hypothesis 3: semantic features are applied more on familiar topics than on unfamiliar topics.

Information quality and trust judgment

Articles varied besides familiarity also in quality. Beyond the hypotheses, the data was used to compare article's quality (determined by WPET) and participants' trust judgments. Both college students ($z = 3.04$, $p < .01$) and PhD students ($z = 2.11$, $p < .05$) did perceive high quality articles as more trustworthy in comparison with low quality articles. By contrast, no significant difference was found for high school students ($z = .80$, $p = .22$) which means that they did not perceive high quality articles as more trustworthy than low quality articles.

4. DISCUSSION

Differences in information feature application

The first hypothesis was partly supported. In line with the hypothesis, PhD students and college students applied more relevant information features (features that were mentioned in the literature) than high school students when judging trustworthiness. However, no difference was found between PhD students and college students in how many and which relevant information features were applied. To discuss these two findings, first the expected difference between the user groups will be reviewed, and after that the unexpected similarity between college students and PhD students will be discussed.

The majority of high school students mentioned five relevant information properties: accuracy, completeness, images, length, and writing style. The college students and PhD students mentioned these five and another three: authority, objectivity, and structure. This increase in mentioning relevant information features is supposedly caused by having more information skills (Brand-Gruwel et al., 2005; Macdonald, Heap, & Mason, 2001). It is assumed that, just like other skills, information skills develop with experience and instruction (Siegler & Alibali, 2005). Education progress is likely to have a positive effect on acquiring information skills. At universities, college students (and PhD students even more) are expected to critically evaluate information, and to use appropriate references in their own writings. As expected, authority turned out to be a prominent information feature for college students and even more for PhD students. In contrast, since high school students do not cite references in their own works (Brand-Gruwel et al., 2005), it is not surprising that they also did not pay attention to references in the WST (predicted by Lucassen & Schraagen, 2010). Objectivity and structure could be features like authority that one learns to attend to over time.

College students and PhD students were expected to differ in feature application, but the results showed that they evaluated trustworthiness quite the same way. A similarity between the two groups is their academic level of education, which possibly is the reason why they approach information quite similarly. It is assumed that the difference in education level and progress is larger between high school students and college students than the difference between college students and PhD students. More comparative research should be carried out to verify this statement.

Where college students and PhD students mostly attended to is authority, this feature was not at all attended to by high school students. Instead, their evaluation was to a large extent based on the accuracy of information. In the current study the inclination to mention accuracy was only found for high school students; the other two groups mentioned accuracy only scarcely. That is, a majority of the college students and a majority of the PhD students mentioned accuracy, but the number of those remarks was far less than for high school students. This finding was unexpected, because all participants were due to the familiarity manipulation on the one hand domain experts toward five Wikipedia articles, and thus likely to

apply accuracy, and on the other hand novices toward five other Wikipedia articles, where they could not apply prior knowledge to verify accuracy of information. According to Scholz-Crane (1998), college students, like high school students, attend to information accuracy. However, this was not found in the current study. This discrepancy in Scholz-Crane's findings and the present findings may be caused by an inconsistency of college students in what they say they will attend to (which was the task in Scholz-Crane, 1998) and how they actually act (which was the task in the current study). It is also possible that source of information (two websites in Scholz-Crane, 1998; Wikipedia in the current study) influenced feature application (see the 3S-model). In that case, two sources make college students to attend more to accuracy than one source. The current study is not sufficient to validate this suggestion.

Differences in strategy application

In hypothesis 2 it was predicted that the three groups differed in strategy application, with PhD students as the group that applied surface strategy the most. Before strategy application is discussed, one should bear in mind that the strategies to make a trust judgment should not be seen separately. When Lucassen & Schraagen (2011) introduced the 3S-model, they pointed out that users apply a combination of source, surface and semantic strategies. The results and interpretations of strategy application should be seen in light of that consideration.

Results indicated that high school students based their trust judgments less on surface strategy than college students and PhD students. However, college students and PhD students did hardly differ in strategy application. An explanation for this absence of effect could be that there is a kind of upper limit on possessing information skills. Apparently, this limit is reached earlier than expected, namely by college students already, while it was expected that PhD students had even more information skills than college students. In reference to Brand-Gruwel et al. (2005), who found differences in trust judging behavior between experts and novices, PhD students and college students could be considered experts against high school students who can be seen as novices. The absence of a noticeable difference in strategy application between college students and PhD students could also be caused by the quantitative method of testing. The coding scheme only counts the number of remarks made by the participants, but it does not provide information about the quality of the evaluation. In other words, college students and PhD students seemingly apply approximately the same amount of surface and semantic features, but further analysis should reveal if this proportion also leads to the same trust judgment.

The reason for high school students to semantically approach information may be that their range of surface skills is not yet developed to such a degree to apply them. At least it is thought that high school students did not intentionally apply a semantic strategy over a surface strategy.

As expected, none of the participants explicitly evaluated source features (i.e. comments about the quality of Wikipedia were absent). Two reasons can be given why nobody commented about the source of information. First, during the experiment participants could not directly compare Wikipedia with other information sources, or as Fogg and Tseng (1999) phrase it, there were “no reference points for comparison.” The second reason is that, after all, the online source was well-known to the user groups, considering the statistics that Wikipedia is used relatively often at school⁶ and that it is frequently used by college students to find background information (Lim, 2009). The fact that all participants were familiar with Wikipedia might have effected trust ratings, since source experience can lead to a biased trust judgment (Lucassen & Schraagen, 2011; in reference to the 3S-model, displayed in Figure 1). Assumedly, source evaluation will emerge when more than one source is presented and the sources are unknown to the user.

Strategy and familiarity

Familiarity with a topic and the application of a semantic strategy were positively related. When users evaluated an article that was familiar to them, they applied semantic strategy more than they did when an article was unfamiliar. This finding is in line with Lucassen and Schraagen (2010), who already found this result for college students. As familiarity was a manipulation of the user characteristic domain expertise, the relation between familiarity and semantic strategy application also goes for domain expertise and semantic strategy application.

High school students and Wikipedia

Now that the hypotheses are discussed, two other findings need further explanation: high school students do not mention that Wikipedia is an open source, and high school students evaluate high and low quality articles equally.

First, high school students did not give complete descriptions of what characterizes Wikipedia. Only 3 out of 13 high school students mentioned the open character of the online encyclopedia. Apparently, this group of students is not aware of the consequences for a source to be open for editing or not. This observation, that high school students do not know important characteristics of a source, can be understood with the ‘Scheme of Student Development’ (Perry, 1970). According to Perry, students go through nine stages, grouped in four areas. The first area is dualism (information is true or false), then comes multiplicity (all information is potentially true), followed by contextual relativism (opinions require support), and the last area is commitment within relativism (different viewpoints of truth can be

⁶ <http://www.alexa.com/siteinfo/wikipedia.org#>

taken). Possibly, the high school students who did not describe Wikipedia as an open source are in the second area, that of multiplicity. They are interested in any source, as long as it contains information, for there is always a possibility that some statements are true. Another reason why high school students neglected the open character of Wikipedia could be that they did not hear about critics like Seigenthaler denouncing the vulnerability of Wikipedia. Because of their age and interests, college students and PhD students are more likely to have heard of criticism towards Wikipedia.

Second, high school students did not differentiate between high and low quality articles in their trust judgments. College students and PhD students rated high quality articles as more trustworthy than low quality articles, but this effect was not observed for high school students. As already discussed, results showed that college students and PhD students distinguished themselves with high school students in their attention to the information features authority, objectivity and structure. These three information features are relevant for information quality; hence paying attention to those features helps in differentiating high quality articles from low quality articles. Another explanation why high school students did not differentiate high versus low quality articles may be that perhaps their attitude toward online information is influenced by a strong so-called ‘willingness to trust’ (Chopra & Wallace, 2003). A higher willingness to trust involves taking a higher risk (Kelton et al., 2008). To take a risk in trusting online information means that there is a chance of making a mistake. For instance, one can trust information while its credibility is low. In that case, the user makes a ‘gullibility error’ (Fogg & Tseng, 1999). It is also possible that the user makes the mistake of not trusting information, while information’s credibility is high (‘incredulity error’). An interpretation of the finding that high school students could not see the difference between a high quality article and a low quality article could be that

they accept a higher risk to wrongfully trust information. High school students are assumed to be less experienced in overseeing the consequences of wrongfully trusting information, whereas college students and PhD students are assumed to have learned that.

Limitations and future research

Earlier research of trust judgment on websites showed that users especially pay attention to design look, structure, and information focus (Fogg et al., 2003). Compared to that, participants in the current study seemingly assess credibility quite differently. However, this can be explained with regard to two differences in experimental design: the number of websites assessed by participants and their familiarity to those websites. In Fogg et al., each participant viewed 10 articles which were unfamiliar to him, while in the current study only one website, which was familiar, was viewed. Therefore it is assumed that presence of other sources and familiarity of the source influence which information features are most mentioned.

When users evaluate information semantically, it was often based on recognition (i.e. to see something in an article and checking this with knowledge from one's long term memory) instead of recall (i.e. to retrieve knowledge from one's long term memory and searching for this in an article). Although it rarely happened that participants expressed an expectation prior to looking at the information, one of the participants said to expect an image of Johan Crujff while looking at an article of FC Barcelona. However, far more often participants said to recognize facts and pictures, thus after looking at the information. No instructions were given that might have influenced participants' tendency to recognize information instead of recall, nor is it likely that the think aloud method affected the cognitive progress (Van Someren et al., 1994). An explanation why participants based their evaluation more on recognition than on recall could be that recognition primes the participant, and priming makes the user to retrieve knowledge from memory "extremely rapidly" and also automatic (Ashcraft, 2006).

The order of components (introduction, table of contents, etc.) was the same for each article, which determines for most of the participants what to read first. This effect is in line with the prominence interpretation theory (Fogg et al., 2003). Participants are guided by the order of an article, meaning that one can speak of a bottom up approach (article driven). However, the current study focused on perception of users towards information (user oriented). In future research a method could be considered that varies the order in which information components are presented to reduce the effect of prominence interpretation. Yet, when the order of information components is changed dramatically, ecological validity will decrease. For example, information pieces never end with an introduction.

The high school students in this study did not have a thorough command of the English language, so that Dutch Wikipedia-articles were presented to them. As earlier described, the English Wikipedia-articles are rated according to a quality rating system (see Table 2); the Dutch articles however are not rated like that. Ideally, in future research only one language edition (preferably English) is used.

Implications of this study are there for information users, teachers, and researchers. To start with, users get the opportunity to reflect their behavior toward trust judgment. Furthermore teachers will be better equipped in understanding trust assessment of their students, and last but not least, researchers can elaborate on the theme of trust judgment. More specifically, findings with regard to trust judgment on Wikipedia (which is the information source assessed in the experiments) may contribute to improving of instruments that try to visualize Wikipedia's credibility (e.g. WikiTrust⁷, WikiDashboard⁸, Wikirage⁹, and WikipediaViz). More generally, insight in how information's credibility is evaluated by different users may inspire everyone who is interested in presenting information. As an illustration, this applies to

⁷ <http://wikitrust.soe.ucsc.edu/>

⁸ <http://wikidashboard.appspot.com/>

⁹ <http://www.wikirage.com/>

journalists who want to write an attractive column, it applies to advertisers who want to send persuasive messages, and it applies to politicians who want to convince their grassroots.

Only one information source was used in the current study. Consequently, the source strategy from the 3S-model was neglected. Further research should focus more on the role that source of information plays in trust judgment.

5. ACKNOWLEDGMENTS

Thanks go to all students who participated in the WST, to Andreas Bremer, Knut Jägersberg and Koen Remmerswaal for their effort in gathering data, and to Teun Lucassen and Jan Maarten Schraagen for supervising the thesis.

6. 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.
- Alexander, J. E. & Tate, M. A. (1999). *Web Wisdom; How to Evaluate and Create Information Quality on the Web*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1st edition.
- Ashcraft, M. H. (2004). *Cognition* (4th edition). Upper Saddle River, NJ: Pearson Education Inc.
- Ballou, D., Wang, R., Pazer, H., & Kumar, G. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*, *44*, 462–484.
- Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*, 271–286.
- 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.
- Cassel, R. (1995). Selection Criteria for Internet Resources. *College and Research Library News*, *56*, 92–93.
- Chevalier, F., Huot, S., & Fekete, J.-D. (2010, March). WikipediaViz: Conveying article quality for casual Wikipedia readers. In S. North, S. Han-Whei, & J. J. van Wijk (Chairs), *IEEE Pacific Visualization Symposium*. Symposium conducted at the meeting of the Electrical and Electronics Engineers, Taipei, Taiwan.
- Chopra, K. & Wallace, W. (2003). Trust in electronic environments. In R. H. Sprague, (Ed.), *Proceedings of the 36th Annual Hawaii International Conference on Systems Sciences: Vol. 9*. Hawaii, USA: IEEE Computer Society Press.
- Crawford, H. (2001). Encyclopedias. In R. Bopp, L. C. Smith (Eds.), *Reference and information services: an introduction* (3th ed., pp. 433–459). Englewood, CO: Libraries Unlimited.
- Dondio, P., Barrett, S., Weber, S., & Seigneur, J. (2006). Extracting trust from domain analysis: A case study on the wikipedia project. In L. Yang, H. Jin, J. Ma, & T. Ungerer, (Eds.), *Autonomic and Trusted Computing: Vol. 4158. Lecture Notes in Computer Science* (pp. 362–373). Berlin: Springer.
- Dondio (2007). Computational trust in web content quality: A comparative evaluation on the Wikipedia project. *Informatica (Ljubljana)*, *31*, 151–160.

- Emigh, W. & Herring, S. C. (2005). Collaborative authoring on the web: A genre analysis of online encyclopedias. In R. H. Sprague (Ed.), *Proceedings of the 38th Annual Hawaii International Conference on System Sciences: Vol. 4*. Washington, DC, USA. IEEE Computer Society Press.
- English, L. P. (2005). To a high IQ! Defining information quality: more than meets the eye. IDQ newsletter 1. Retrieved April 29, 2011, from: <http://iaidq.org/publications/doc2/english-2005-04.shtml#english4>.
- Eppler, M. & Muenzenmayer, P. (2002). Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In C. Fisher & Davidson, B. N. (Eds.), *Proceedings of 7th International Conference on Information Quality* (pp. 187–196).
- Fogg, B.J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences* (pp. 1–15). New York: ACM.
- Fogg, B.J. & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit* (pp. 80–87). New York: ACM.
- Fritch, J. W., & Cromwell, R. L. (2001). Evaluating internet resources: Identity, affiliation, and cognitive authority in a networked world. *Journal of the American Society for Information Science and Technology*, 52, 499–507.
- Giddens, A. (1990). *The consequences of modernity*. Stanford: Stanford University Press.
- Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438, 900–901.
- Grimes, D. J. & Boening, C. H. (2001). Worries with the web: A look at student use of web resources. *College & research libraries*, 62, 11–23.
- Hovland, C. I., Janis, I. L., & Kelley, J. J. (1953). *Communication and persuasion*. New Haven, CT: Yale University Press.
- Johnson-George, C. & Swap, W. (1982). Measurement of specific interpersonal trust: construction and validation of a scale to access trust in a specific other. *Journal of Personality and Social Psychology* 43, 1306–1317.

- Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59, 363–374.
- Lih, A. (2004). Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism* (pp. 16–17).
- Lim, S. (2009). How and why do college students use wikipedia? *Journal of the American Society for Information Science & Technology*, 60, 2189–2202.
- Lorenzen, M. (2002). The land of confusion? High school students and their use of the World Wide Web for research. *Research strategies*, 18, 151–163.
- Lucassen, T., & Schraagen, J. M. (2010). Trust in Wikipedia: how users trust information from an unknown source. In *WICOW '10: Proceedings of the 4th workshop on Information credibility* (pp. 19–26). New York: ACM.
- Lucassen, T., & Schraagen, J. M. (2011). Factual Accuracy and Trust in Information: The Role of Expertise. *Journal of the American Society for Information Science & Technology*, 62, 1232–1242.
- Macdonald, J., Heap, N., & Mason, N. (2001). Have I learnt it?: Evaluating skills for resource-based study using electronic resources. *British Journal of Educational Technology*, 32, 419–433.
- Menchen-Trevino, E. & Hargittai, E. (2011). Young adults' credibility assessment of wikipedia. *Information, Communication & Society*, 14, 24–51.
- 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.
- Motro, A., & Rakov, I. (1998). Estimating the quality of databases. In *Proceedings of FQAS 98: Third International Conference on Flexible Query Answering Systems* (pp. 298–307). Roskilde, Denmark.
- Naumann, F., & Rolker, C. (2000). Assessment methods for information quality criteria. In *Proceedings of the International Conference on Information Quality* (pp. 148–162).
- O'Keefe, D. J. (2002). *Persuasion: Theory and research* (2nd ed.). Thousand Oaks, CA: Sage.

- Petty, R., Cacioppo, J., & Goldman, R. (1981). Personal Involvement as a Determinant of Argument-Based Persuasion. *Journal of Personality and Social Psychology*, 41 847–855.
- Scholz-Crane, A. (1998). Evaluating the future: A preliminary study of the process of how undergraduate students evaluate Web sources. *Reference Services Review*, 26, 53–60.
- Seigenthaler, J. (2005). A false Wikipedia ‘biography’. Retrieved April 28, 2011, from: http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm.
- Siegler, R. S., & Alibali, M. W. (2005). *Children’s Thinking* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Stanford Web Credibility Guidelines. Retrieved November 14, 2011, from: <http://credibility.stanford.edu/guidelines/>.
- Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the 2005 International Conference on Information Quality*.
- Van Someren, M. W., Barnard, Y. F., & Sandberth, J. A. C. (1994). *The think aloud method: a practical guide to modelling cognitive processes*. London: Academic Press.
- Viégas, F., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of SIGCHI* (pp. 575–582). Vienna, Austria.
- Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Computers & Education*, 52(1), 234–246.
- Walthen, 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.
- Wang, R. Y., Allen, T. J., Harris, W., & Madnick, S. E. (2003). An information product approach for total information awareness. *SSRN Electronic Journal*.
- Wikipedia Featured Article criteria: Retrieved November 14, 2011, from: http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

- Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Westport, CT: Greenwood Press.
- Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, 55, 249-270.
- Yang, H.-L. & Lai, C.-Y. (2010). Motivations of wikipedia content contributors. *Computers in Human Behavior*, 26, 1377–1383.
- Zhu, X. & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Research and Development in Information Retrieval* (pp. 288–295).

7. APPENDIXES

Appendix A: Coding scheme for high school students

All participants (high school students)		Articles 1-10													
Component	Introduction		Text		Table of contents		Images		Internal links		References		Other		
	2.7%		80.5%		0.4%		10.9%		1.7%		0.5%		3.2%		
Strategy	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	
	0.9%	1.9%	31.5%	49.0%	0.4%	0.0%	5.6%	5.3%	1.7%	0.0%	0.5%	0.1%	2.2%	1.0%	
Features	General	Completeness	General	Completeness	General	Completeness	General	Caption	General	Quality	General	Completeness	Audio/Video	Audio/Video	
	0.4%	0.0%	6.8%	4.2%	0.4%	0.0%	0.8%	0.5%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	
	Length	Factual accuracy	Length	Currency	Length	Scope	Number	Presence	Number	Presence	Number	Currency	Infobox	Infobox	
	0.0%	1.8%	9.8%	1.4%	0.0%	0.0%	0.2%	0.0%	1.1%	0.0%	0.5%	0.0%	0.1%	0.3%	
	Presence	Objectivity	Quality	Factual accuracy	Presence	Quality	Presence	Quality	Presence	Quality	Presence	Quality	List	List	
	0.0%	0.0%	9.9%	36.4%	0.0%	0.0%	3.0%	2.9%	0.5%	0.0%	0.1%	0.8%	0.0%	0.0%	
	Quality	Quality	Scope	Objectivity	Quality	Quality	Quality	Scope	Quality	Scope	Quality	Scope	Names	Names	
	0.5%	0.0%	0.1%	0.6%	0.0%	0.0%	1.6%	0.7%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	
	Structure	Scope	Structure	Quality	Structure	Quality	Structure	Understandability	Structure	Understandability	Statistics	Statistics	Statistics	Statistics	
	0.0%	0.0%	2.0%	0.8%	0.0%	0.0%	0.0%	1.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	Writing style	Understandability	Writing Style	Scope	Writing Style	Understandability	Writing Style	Understandability	Writing Style	Understandability	Table	Table	Table	Table	
0.0%	0.1%	2.9%	2.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%		
			Understandability										Timeline	Timeline	
			3.2%										0.2%	0.2%	
													Title	Title	
													0.3%	0.0%	
													Topic	Topic	
													0.2%	0.2%	

Figure A1. Coding scheme for all high school students. Most mentioned properties are factual accuracy of text (38.2 %), quality of text (11.2 %) and images (10.9 %). Authority (references; 0.5 %) and objectivity (0.6 %) are less mentioned.

Appendix B: Coding scheme for college students

All participants (college students)		Articles 1-10													
Component	Introduction		Text		Table of contents		Images		Internal links		References		Other		
	4.3%		43.2%		3.4%		12.5%		2.9%		28.4%		5.2%		
Strategy	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	
	2.8%	1.5%	19.8%	23.4%	3.0%	0.4%	7.7%	4.8%	2.8%	0.1%	26.0%	2.4%	5.0%	0.2%	
Features	General	Completeness	General	Completeness	General	Completeness	General	Caption	General	Quality	General	Completeness	Audio/Video	Audio/Video	
	1.6%	0.0%	3.3%	2.4%	0.8%	0.3%	0.8%	0.7%	0.1%	0.1%	9.5%	0.3%	0.1%	0.0%	
	Length	Factual accuracy	Length	Currency	Length	Scope	Number	Presence	Number	Presence	Number	Currency	Infobox	Infobox	
	0.6%	0.7%	7.9%	0.3%	0.5%	0.1%	1.1%	0.2%	0.9%	0.9%	10.1%	0.1%	0.8%	0.1%	
	Presence	Neutrality	Quality	Factual accuracy	Presence	Quality	Presence	Quality	Presence	Quality	Presence	Quality	List	List	
	0.1%	0.1%	3.1%	12.1%	0.7%	4.5%	2.4%	1.8%	5.4%	1.3%	5.4%	1.3%	0.5%	0.1%	
	Quality	Quality	Scope	Neutrality	Quality	Scope	Quality	Scope	Quality	Scope	Quality	Scope	Names	Names	
	0.5%	0.1%	0.4%	1.3%	0.4%	1.2%	0.9%	0.9%	0.0%	0.7%	0.0%	0.7%	0.8%	0.1%	
	Structure	Scope	Structure	Quality	Structure	Understandability	Structure	Understandability	Structure	Understandability	Statistics	Statistics	Statistics	Statistics	
	0.2%	0.2%	3.2%	0.9%	0.6%	0.6%	0.6%	0.6%	1.7%	0.0%	1.7%	0.0%	1.7%	0.0%	
	Writing style	Understandability	Writing Style	Scope	Writing Style	Scope	Writing Style	Scope	Table	Table	Table	Table	Table	Table	
	0.0%	0.6%	1.9%	3.5%	0.0%	0.0%	0.0%	0.0%	0.7%	0.0%	0.7%	0.0%	0.7%	0.0%	
				Understandability					Timeline	Timeline	Timeline	Timeline	Timeline	Timeline	
				3.0%					0.2%	0.0%	0.2%	0.0%	0.2%	0.0%	
								Title	Title	Title	Title	Title	Title		
								0.1%	0.0%	0.1%	0.0%	0.1%	0.0%		
								Topic	Topic	Topic	Topic	Topic	Topic		
								0.1%	0.0%	0.1%	0.0%	0.1%	0.0%		

Figure B1. Coding scheme for all college school students. Most mentioned properties are authority (references; 28.4 %), factual accuracy of text (12.8 %) and images (12.5 %). Less mentioned are currency of text (0.3 %) and presence of table of contents (0.7 %).

Appendix C: Coding scheme for PhD students

All participants (PhD students)		Articles 1-10													
Component	Introduction		Text		Table of contents		Images		Internal links		References		Other		
	9.8%		40.4%		3.7%		6.5%		2.7%		33.4%		3.4%		
Strategy	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	Surface	Semantics	
	6.8%	3.0%	18.9%	21.5%	3.0%	0.7%	4.7%	1.9%	2.7%	0.1%	24.4%	9.0%	3.2%	0.2%	
Features	General	Completeness	General	Completeness	General	Completeness	General	Caption	General	Quality	General	Completeness	Audio/Video	Audio/Video	
	5.5%	0.1%	3.6%	3.1%	1.9%	0.1%	1.2%	0.5%	0.1%	0.1%	7.5%	0.6%	0.5%	0.0%	
	Length	Factual accuracy	Length	Currency	Length	Scope	Number	Presence	Number	Presence	Number	Currency	Infobox	Infobox	
	0.9%	2.0%	8.8%	0.6%	0.0%	0.6%	0.4%	0.1%	1.1%	0.1%	6.6%	1.1%	0.4%	0.0%	
	Presence	Neutrality	Quality	Factual accuracy	Presence	Quality	Presence	Quality	Presence	Quality	Presence	Quality	List	List	
	0.0%	0.1%	1.8%	10.6%	0.8%	0.8%	1.8%	1.3%	1.5%	1.5%	7.2%	6.8%	0.5%	0.1%	
	Quality	Quality	Scope	Neutrality	Quality	Scope	Quality	Scope	Quality	Scope	Quality	Scope	Names	Names	
	0.2%	0.2%	0.9%	1.2%	0.1%	0.1%	1.3%	0.0%	0.0%	0.0%	0.0%	0.4%	0.1%	0.1%	
	Structure	Scope	Structure	Quality	Structure	Quality	Structure	Understandability	Structure	Understandability	Structure	Understandability	Statistics	Statistics	
	0.0%	0.5%	1.9%	0.7%	0.2%	0.2%	0.2%	0.1%	0.2%	0.1%	0.2%	0.1%	1.0%	0.0%	
	Writing style	Understandability	Writing Style	Scope	Writing Style	Scope	Writing Style	Understandability	Writing Style	Understandability	Writing Style	Understandability	Table	Table	
	0.3%	0.1%	1.8%	3.0%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.1%	0.0%	
				Understandability		Understandability		Understandability		Understandability		Understandability	Timeline	Timeline	
				2.3%		2.3%		2.3%		2.3%		2.3%	0.2%	0.0%	
												Title	Title		
												0.0%	0.0%		
												Topic	Topic		
												0.3%	0.0%		

Figure C1. Coding scheme for all PhD students. Most mentioned properties are authority (references; 33.4 %) and factual accuracy of text (12.6 %). Less mentioned are currency of text (0.6 %) and presence of table of contents (0.8 %).

