ABSTRACT
As the use of collaborative online encyclopedias such as Wikipedia grows, so does the need for research on how users evaluate its credibility. In this paper we compare three experimental approaches to study trust in Wikipedia, namely think aloud, eye-tracking, and online questionnaires. The advantages and disadvantages of each method are discussed. We conclude that it is best to use multiple methods when researching information trust, as each single one of the discussed methods alone does not give all possible information.

Keywords
Trust, credibility, Wikipedia, think aloud, eye-tracking, online questionnaires.

INTRODUCTION
Over the last decade, Wikipedia has evolved from a small spin-off project to a massive online encyclopedia, covering almost every imaginable topic. Wikipedia owes its rapid growth to its collaborative nature, with thousands of volunteers contributing to the articles. Information quality has been shown quite high [3]; however, the open-source character also brings a large disadvantage as the authors of the information are mostly unknown. This means that the user can never be sure of the trustworthiness and expertise of the author and thus has to be aware of the possibility of less credible information.

In this paper, we describe three experiments in which the way users cope with the uncertainty of the credibility of information is investigated. First, the Wikipedia Screening Task is introduced. After this, three experiments in which different methodologies were applied are shortly discussed, each of them featuring this task in a different setting.

WIKIPEDIA SCREENING TASK
The Wikipedia Screening Task was first introduced in [4]. In this task, a Wikipedia article is presented, in which obvious cues of credibility, such [citation needed] remarks, are removed. The participant is asked to evaluate the credibility of the article. It is not specified how to perform this task, so the participant is free to incorporate features from the article which he or she deems relevant to credibility.

Various manipulations can be made in experiments featuring the Wikipedia Screening Task. An example is varying the quality of the articles, based on the ratings given by the Wikipedia Editorial Team[1]. A second useful manipulation is the familiarity of the participant with the topic at hand, as different behavior can be expected when a participant has some knowledge on the content.

THINK ALOUD
In this experiment, the Wikipedia Screening Task was performed while thinking aloud [4]. Using this method, the participant is asked to verbalize everything that comes to mind while performing a task [2]. Audio is recorded and typed out afterwards. Based on this, the utterances of each participant can be categorized using a coding scheme. Using only a few participants (N=12), a comprehensive list of features relevant to trust could be established. It was found that the most important features were references (quantity and quality), several textual features (e.g., length, comprehensiveness), and pictures (e.g., quality, relevance). Article quality and familiarity with the topic were manipulated. The provided credibility ratings indicated that the participants (college students) were able to distinguish good and poor information quality. However, no differences in the features used could be found between good and poor information quality as well as familiar and unfamiliar topics.

Think aloud is a great method to gather a lot of rich information on the behavior of participants. Direct insights into cognitive processes of the participants are gained, as data is gathered during task execution, instead of afterwards. Only a few participants were needed to form a comprehensive list of features relevant to trust.

On the downside, the data analysis of think-aloud is very labor-intensive. All utterances of each participant have to be fully typed out before they can be categorized. Furthermore, two coders are needed in order to calculate inter-rate reliability. A second drawback is that thinking-aloud does not come naturally to everybody. Some participants will be more capable of expressing their thoughts than others.

EYE-TRACKING
In a second experiment, the Wikipedia Screening Task was performed while the gaze of the participants was constantly monitored using an eye-tracker [6]. Gaze is a very good indicator of visual attention [1]. In this experiment, we were particularly interested in the use of references in credibility evaluation. Four strategies of reference use in credibility evaluation were proposed: (1) references are not considered, (2) the presence of references is checked, (3) the number of references is checked, and (4) the quality of references is checked.

Each of these strategies can easily be distinguished by visual inspection of the eye-tracking data. It was found that each of the strategies was applied, with the dominant strategy being the first, namely not to consider the references at all.

This observation seems to contradict to the findings in the think aloud experiment, in which references were found to be very important to the participants. This may indicate socially desirable behavior in the first experiment. However, the lack of visual attention does not necessarily mean that references were not part of the mental model of the participants.

Eye-tracking has the advantage to be less obtrusive than think aloud. Using a table-mounted device, participants are somewhat restricted in their movements, but they are not required to perform a secondary task. Moreover, gaze is a very good indicator of visual attention, so the elements of an article which are attended when evaluating credibility can be assumed to be of relevance to the participants.

A drawback of eye-tracking is that while we know what objects are attended, we don’t know why. One could suggest a combination of eye-tracking and think aloud to obtain this information. However, this is not advisable since think aloud tends to slow down task performance, which may lead to different attention. A possible solution is retrospective think aloud. In this method, the participant first performs the task without thinking aloud (but possibly with an eye-tracker). Video is recorded and played back directly afterwards. During the play-back the participants is asked to verbalize what they were thinking when they were performing the task.

ONLINE QUESTIONNAIRES
Whereas the think aloud and eye-tracking studies both took place in a controlled lab-environment, a different approach was taken in a third experiment, which was performed online [5]. Over 650 Internet users took part in a short experiment, in which the influence of factual accuracy on trust of experts and novices was examined. This was done by showing them articles on car engines, featuring varying levels of deliberate errors.

It was shown that the experts were influenced by the errors, whereas the novices were not. Moreover, proof was found for the proposed 3S-model, in which three strategies were proposed in which domain expertise, information skills, and source experience lead to different features in the information being noticed.

Obviously the biggest advantage of online questionnaires is that a high number of participants can easily be recruited, for instance by posting requests for participation on online forums. This also means that participants with particular characteristics can be addressed, for instance automotive experts.

However, these advantages come with an immediate drawback: the behavior of the participants can hardly be accounted for. It can never be assured that each participant takes the experiment seriously. Therefore, the data should always be inspected manually for bogus answers. Furthermore, it is very hard to repeat an experiment or perform a similar experiment, as the same participants (with unwanted prior knowledge) may participate again.

CONCLUSION
In this paper we showed three methods to combine with the Wikipedia Screening Task. Each method has its own advantages and disadvantages, but the most important lesson learned is that different approaches yield different results. When researching online trust, multiple methods should be apprehended to avoid that conclusions on credibility evaluation behavior are biased by the methodology applied.

REFERENCES