


Information quality assessment of community-generated content – A user study of Wikipedia

Journal of Information Science
37(5) 487–498
© The Author(s) 2011
Reprints and permission: sagepub.
co.uk/journalsPermissions.nav
DOI: 10.1177/0165551111416065
jis.sagepub.com


Eti Yaari

Department of Information Science, Bar-Ilan University, Israel

Shifra Baruchson-Arbib

Department of Information Science, Bar-Ilan University, Israel

Judit Bar-Ilan

Department of Information Science, Bar-Ilan University, Israel

Abstract

This study examines the ways in which information consumers evaluate the quality of content in a collaborative-writing environment, in this case Wikipedia. Sixty-four users were asked to assess the quality of five articles from the Hebrew Wikipedia, to indicate the highest- and lowest-quality article of the five and explain their choices. Participants viewed both the article page, and the article's history page, so that their decision was based both on the article's current content and on its development. The analysis shows that the attributes that most frequently assisted the users in deciding about the quality of the items were not unique to Wikipedia: attributes such as amount of information, satisfaction with content and external links were mentioned frequently, as with other information quality studies on the web. The findings also support the claim that quality is a subjective concept which depends on the user's unique point of view. Attributes such as number of edits and number of unique editors received two contradictory meanings – both few edits/editors and many edits/editors were mentioned as attributes of high-quality articles.

Keywords

Wikipedia; user-study; information quality assessment

1. Introduction

Information literacy is one of the basic skills in the twenty-first century. The American Library Association (ALA) stated in 1989 [1]:

To be information literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate and use effectively the needed information.

Thus critical evaluation of information is of utmost importance. In the print world, before publication, materials undergo reviewing processes, and information production is controlled by publishers and by information providers whose reputation rests on their professional standing [2]. This skill becomes even more essential on the web, where almost anyone can publish almost any information almost without any restrictions. Thus the 'burden' of examination, as well as the control over the quality of information has shifted from the experts to the information-seeking public [3, 4].

Corresponding author:

Judit Bar-Ilan, Department of Information Science, Bar-Ilan University, Ramat Gan, 52900, Israel.
Email: barilaj@mail.biu.ac.il

There is no generally accepted definition of information quality, as discussed by Arazy and Kopak [5]. According to Hilligoss and Rieh [6]: 'Information quality refers to people's subjective judgment of goodness and usefulness of information in certain information use settings with respect to their own expectations of information or in regard to other information available' (p. 1469). According to Rieh [4] information quality comprises five facets: usefulness, goodness, accuracy, currency and importance. Other related concepts are credibility [4, 7], and cognitive authority of the information [4]. One of the frequently cited criteria, mentioned by users when evaluation information on the web, is its authority, i.e. the reputation of the source or of the author (e.g. [4, 7–10]). It is interesting to note that Eysenbach and Köhler [9] found that, although information consumers state that they verify the reliability of on-line information, including checking the sources' authority, in practice they do not necessarily do so. Additional criteria mentioned in information evaluation studies include list of links, site design [10, 11], information organization [11] and site navigation [11]. Rieh and Danielson [12] provide an extensive review of the concept of information quality.

The ELM (the Elaboration Likelihood Model) of Petty and Cacioppo [13] is often used to explain some of the findings in evaluation studies (e.g. [2, 10]). According to the ELM, users take the 'central route' when they are motivated and have knowledge about the topic on hand, and take the 'peripheral route' when motivation and knowledge are lacking, and make decisions based on external features like the design of the site.

Wikipedia is one of the most popular sites on the web. According to Alexa [14] it is currently (as of June 2011) the seventh most visited site on the web, and the top information source, as the sites above it on the list are either search engines or social network sites. Thus it is of utmost importance to study how users evaluate information quality of Wikipedia. As can be seen from the above discussion, the reputation of the author and the source of the information is a major criterion for evaluation information on the internet in general. This criterion is not suitable for evaluating Wikipedia articles, because the author of the information in Wikipedia is not defined. Wikipedia articles are the results of combined efforts by Wikipedia contributors, and everyone can become such a contributor.

There is an ongoing discussion about the value and reliability of information produced by the crowd (the concept of the 'wisdom of crowds' [15]) versus information produced by experts. An early study comparing the quality of Wikipedia with *Encyclopaedia Britannica* [16] indicated that the quality of Wikipedia entries was similar to that of the Britannica entries. A different conclusion was reached by Rector [17], who found that, based on nine articles, Wikipedia's accuracy was considerably lower than that of the Britannica, the Dictionary of American History or American National Biography Online.

Several measures have been proposed to evaluate the quality of Wikipedia articles. Some of these measures are based on the history of the article: the number of revisions or the number of editors involved. Lih [18] suggested evaluating the articles based on the article history (total number of edits and total number of unique editors). He considered Wikipedia articles cited in English language press. Zeng et al. [19] defined trustworthiness based on revision history. They evaluated their model on English language articles from the geography category, and showed that their model was able to differentiate between featured, cleanup and normal articles. 'Featured articles are considered to be the best articles in Wikipedia, as determined by Wikipedia's editors' [20]. Articles receive the featured attribute through community voting. Cleanup articles are those deemed to be in need of major revision and are marked by Wikipedia editors. Adler and de Alfaro [21] proposed to define Wikipedia editor/author reputation based on the length of time the text edited remains 'alive'. They analysed all articles in the Italian and the French Wikipedia and showed that the texts of editors with lower reputations are shorter lived. In a later paper Adler et al. [22] devised a method where trust depended on the reputation of the authors and the reputation of the editors who subsequently revised the entry. This time they validated the results based on the English Wikipedia and showed that text labelled as 'low trust' had significantly higher probability of being edited than text labelled as 'high trust'. Hu et al. [23] also devised measures based on the revision history more precisely on the interaction between articles and their contributors. They tested the measures on Wikipedia articles on countries, and used Wikipedia's grading scheme for the articles. This scheme includes the 'featured articles' mentioned earlier along with five additional grades, as assigned by the Wikipedia: Version 1.0 Editorial Team for the English Wikipedia [24]. Wilkinson and Huberman [25] were able to distinguish between featured and regular articles in the English language Wikipedia based on the number of edits, editors and the intensity of the cooperation behaviour. Note that Kittur and Kraut [26] studied the relations between quality and the number of editors involved in creating the entry, and found that increasing the number of editors improved quality only when the editors coordinated their work. For evaluation they also used the labelling of the Wikipedia Editorial Team Assessment Project [24]. Wöhner and Peters [27] based their quality assessment on the persistence of the contributions for Wikipedia articles, and showed that for the German Wikipedia these metrics were able to differentiate between high-quality (featured articles) and low-quality articles. Stein and Hess [28] claimed that it mattered who contributed, and measured this by the number of pages an author edited and by the total number of the author's edits in the German Wikipedia.

Stvilia et al. [29] developed seven measures that were able to discriminate well between random and featured Wikipedia articles from the English language Wikipedia. These measures were: authority/reputation – based on the number of editors, edits, reverts, external links; completeness – based on the number of broken links and article length; readability; informativeness; consistency; currency; and volatility. Dondio, Barrett, Weber and Seigneur [30] also suggested a complex measure in order to estimate the trustworthiness of Wikipedia articles. The measure was evaluated on a set of 8000 English Wikipedia articles. Blumenstock [31] recommended a simplistic approach to evaluate the quality of Wikipedia articles: the number of words in the entries. His observation was based on the distribution of the word length of featured versus random articles in the English Wikipedia.

As can be seen from the above review, many measures have been proposed to evaluate the quality of Wikipedia articles, but we were only able to find a few papers where users made quality assessments. Kittur, Chi and Suh [32] tested the effects of visualization of article history. Users recruited through Amazon's Mechanical Turk were asked to assess the trustworthiness of Wikipedia articles on a seven-point scale. The results showed that the suggested visualization had an impact on perceived trustworthiness. Wikidashboard (<http://wikidashboard.appspot.com/>) provides a visualization of article and author editing history for entries and editors in the English language Wikipedia and in Wikipedias in several other languages. In an experiment conducted by Pirolli, Wollny and Suh [33], 24 users evaluated the credibility of six articles, half with the Wikidashboard interface and half without. The users who saw the dashboard assigned greater credibility to the articles. Chesney [34] asked academics to read a Wikipedia article and to assess its credibility. Half the participants were asked to read an article within their area of expertise and half were given a random article. After reading the article the participants were asked to fill in a questionnaire with Likert-scale questions on credibility. He found that the users assigned higher credibility to the article that was within their knowledge domain. Nofrina et al. [35] used focus groups to elicit criteria used attributes of wikis that can serve as credibility cues. These attributes included: editing and history features, presence of references, links to other sites and comments in the discussion section ('talk pages'). Focus groups were also used by Metzger, Flanagan and Medders [36] in order to study how users evaluate credibility of information and whether they view information provided by social computing applications as credible. Findings from this study showed that users often rely on others to make credibility assessments and they often invoke cognitive heuristics (rules-of-thumb, guidelines) for evaluation instead of systematic processing of information.

As can be seen from the above review, there are only a very few user studies where readers were asked to assess the quality of Wikipedia articles. We could not find any previous study where users were observed while evaluating the quality of Wikipedia articles. Such observations, using a qualitative methodology, allow characterization of the criteria employed by users. The users are not influenced by criteria suggested by the researchers neither through specific interfaces nor through questionnaires. In this paper we report on an extensive user study, where our objective was to try to understand how users perceive information quality when the information comes from Wikipedia. In particular we asked the following research questions:

1. Which criteria assist users in determining that a Wikipedia article is of high quality?
2. Which criteria assist users in determining that a Wikipedia article is of low quality?

Two separate questions were phrased, because it is quite plausible that different criteria are being employed for deciding that an article is of high quality and that it is of low quality. In addition, since many previous studies relied on community labelling of Wikipedia articles, we asked:

3. Do users consider featured articles as high-quality articles?

Since the population of the study was Hebrew speaking, articles from the Hebrew Wikipedia were selected. The Hebrew Wikipedia was established in 2003 and as of June 2011 it comprises more than 115,000 articles [37]. According to the Hebrew Wikipedia statistics page, in September 2010 there were 34,384,220 page views [38], and according to Sue Gardner, executive director of the Wikimedia Foundation, in May 2009, 1.4 million users used the Hebrew Wikipedia per month [39].

2. Research setup and methods

The study employed a qualitative method: users were observed during the evaluation process. They were asked to 'think aloud', and they were interviewed about their perceptions. The users' task was to assess five Wikipedia articles, based on their content and history page, and to pick the articles they considered 'best' and 'worst'. They were asked to justify their choices. 'Best' and 'worst' were subjectively interpreted by the users.

2.1. Population and sample

The 64 participants in this study were undergraduate and graduate students from Bar-Ilan University, 52 of them from the Department of Information Science, seven from the Faculty of Exact Sciences and five from the Faculty of Life Science. This population uses Wikipedia frequently both for studying and for everyday purposes [40]. Information Science students were rewarded with bonus points in their respective classes, while students from other departments received a small monetary reward for their participation.

Out of the 64 participants, 19 (29.7%) were men and 45 (70.3%) women; the participants' average age was 30.6 years ($SD = 7.93$; median = 28). The youngest participant was 22 years old, and the oldest 51. Undergraduate students constituted 39.0% ($n = 25$) of the study population, and 61.0% ($n = 39$) were graduate students. Most participants ($n = 56$, 87.5%) had been aware, for at least one year, of the existence of Wikipedia, with many ($n = 30$, 46.9%) reporting having known about it for more than two years. The participants were aware of the existence of Wikipedia for 1.6 years on the average ($STD = 0.61$). The same percentage (46.9%, $n = 30$) reported that they had visited Wikipedia following a link from a search engine result page. A considerable number of the users (40.6%, $n = 26$) visited Wikipedia at least once a week, and only eight participants (12.5%) had not used it before. Only six participants (9.4%) had edited a Wikipedia article.

2.2. Article selection

Wikipedia community assigns templates to articles as a result of a community decision. These templates are called 'Wikipedia templates'. The 'featured articles' are considered to be the best articles by Wikipedia editors. A 'needs expansion' template is added to articles which do not have enough content, and there are also articles marked as 'requiring cleanup (editing)' or 'needing rewrite'. The templates, except for the 'featured article' template, hint that there might be some quality issues with articles. The templates, except for the 'featured article' template can be added by any Wikipedia editor. There are of course articles with no Wikipedia templates attached to them – we call these 'regular articles'. In our study we wanted our participants to evaluate articles from each group.

We randomly picked six major Wikipedia categories out of the 11 major categories from the Hebrew Wikipedia: arts, history, life sciences, literature, religion and technology. Articles were randomly sampled from each category and from this sample a final selection was made in such a way that for each category and for each of the four Wikipedia templates (featured, expand, cleanup and rewrite) five articles were selected. Thus the selection process resulted in 120 articles (six categories, 20 articles in each category). In addition a single 'regular article' was selected for each category. The selected articles were downloaded and saved together with their history pages, so that the participants would all see stable versions of these articles. The history pages were also saved and shown to the users, because a large number of the previously proposed quality measures rely on the version history of the article, and we assumed that seeing the revision history might influence the users' judgements. Although saved versions meant that participants could not see any subsequent updates, it was decided that it was more important that the users saw identical versions of the articles and their history pages. There was no noticeable effect on the availability of external links or on the updatedness of the articles.

2.3. The task

Each participant had a one-on-one meeting with the first author. First the purpose of this study was described – trying to understand how users evaluate information from the Hebrew Wikipedia – and then participants were told about what they were expected to do during the meeting. They signed an informed consent form, agreeing also to an audio recording of the meeting.

Each participant was asked to choose one of the six content categories, and then was presented with four lists of five article names each and was asked to pick one article from each list. The lists corresponded to the Wikipedia templates, but this information was not provided to the participants. In addition the interviewer asked them to evaluate a fifth article (this was the 'regular article').

The saved versions of the selected articles were opened each in a separate browser window. The participants were instructed to view the five screens and were allowed to go back and forth if they wanted to compare them. The participants were also shown the saved history tab of each article, showing its revision history. Half of the participants were first shown the article then its revision history, and half were first shown the revision history and then the article itself. The participants were given detailed explanations regarding the meaning of the revision history page. Half of the participants saw the content pages together with their Wikipedia templates, and the other half saw the pages with the templates removed.

The participants were instructed to 'think aloud' while they viewed the articles and their revision histories. They were asked to indicate which of the five articles was of the lowest and the highest quality and were asked to explain their choices.

Following the evaluation of the articles by the users, they were debriefed, the aim of the study was presented in depth and any questions the users had regarding the study were answered. The average meeting lasted 58 minutes (SD = 3.3, median = 59); the shortest meeting lasted 55 minutes and the longest 70 minutes.

2.4. Data analysis

The verbal data collected were fully recorded and transcribed, and saved as separate Word files for each participant. Following the procedures for qualitative research, the data were analysed using content-analysis techniques, unitizing and classifying these units by context [41]. The unit of analysis was defined here as any verbal statement made by the participant in a certain context, regardless of its length [42]. In the first stage, each transcript of each participant was read on its own and analysed, while notes were taken and an initial system of categories was designed. During this stage we took great care to maintain coding independence, namely, that the analysis of a given transcript would not influence the reading of subsequent ones. The second stage consisted of an examination of the previously identified initial categories, comparing them, and placing them on horizontal and vertical axes to form a category tree [41]. This procedure was repeated several times, as each new category identified led to a reexamination of all other categories to verify that no unit of analysis required reclassification. Following this, the third stage consisted of defining the main categories, and the final category tree was established.

The analysis carefully followed the rules of qualitative research and several steps were taken to ensure its validity: (1) the classification of the units and category placement on the axes were discussed with an additional judge; (2) prior to establishing the final category tree, eight random transcripts were coded by a different judge as a control for the classification process (the intercoder reliability was 89%); and (3) the analysed material (transcripts, comments, and all stages of coding and mapping) was saved in full as evidence for the analysis procedure.

3. Results

3.1. Criteria for assessing Wikipedia articles

The analysis of the participants' arguments for choosing the highest and lowest quality items from the five articles they viewed revealed that the criteria mentioned can be divided into two major categories: *non-measurable* and *measurable*.

By *measurable* we mean criteria that can be objectively and reliably assigned by a computer program without human intervention (e.g. the number of words in an article or the existence of images). The *non-measurable* criteria were more subjectively assigned by the users (e.g. structure, relevance of links, writing style). Arazy and Kopak [5] in a recent paper discuss the measurability of information quality. We further subdivided the criteria into criteria mentioned in relation to the content page and in relation to the history page of the articles. The categorization appears in Figures 1 and 2. In the research questions we differentiated between criteria for identifying high- and low-quality articles, but it turned out that the same criteria were used to identify both, where for high-quality articles the criterion was used in a positive sense and for low-quality articles in a negative sense. In Tables 1 and 2 the distributions of the comments for high- and low-quality articles are displayed.

The most frequently mentioned non-measurable criterion both for high- and low-quality articles was the article's coverage and scope (23 positive and 17 negative remarks from the 64 participants; mentioned by 35.94% and 26.56% of the participants respectively), followed by comments on its structure (12 positive and five negative remarks made by 18.75% and 7.81% of the participants respectively). Writing style was also a criterion, as one of the participants stated: 'The article is explained really well I can give it to my wife, who probably has not heard the term, I can give it to her and I know she'll understand'. An additional criterion was the relevance of internal links (mentioned by three participants (4.69%) for low-quality and three participants (4.69%) for high-quality articles). By 'internal links' we mean links within Wikipedia, while external links refer to non-Wikipedia sources on the internet, and references to print sources. Here is an example of a comment on a low-quality article: 'These links seemed very amateurish to me. There should be internal links only to the important concepts, the ones that really need explaining'. The participants made only a very few comments related to external link quality (mentioned by one participant (1.56%) for low-quality and three participants (4.69%) for high-quality articles).

The non-measurable criteria mentioned when viewing the revision history page were rather vague and few; some participants mentioned the quality of the comments (positive remarks by 14, and negative remarks by four users; 21.88% and 6.25% of the participants respectively). There were a few interesting remarks on the nicknames Wikipedia editors used; for example here is a negative comment: 'The nicknames give me the feeling that it is not all that serious, that the article is not serious, that it is turned into some sort of entertainment'. Two participants commented on the names and nicknames of the editors for the article identified by them as high quality (2.23%) and two participants (2.23%) made such comments for the low-quality article.

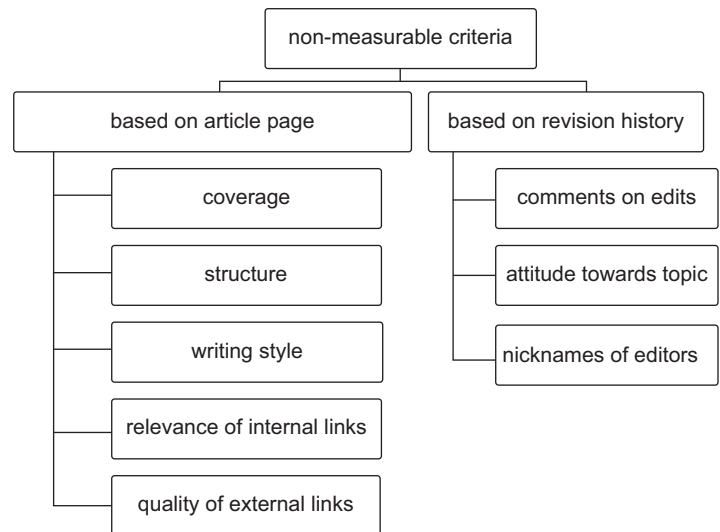


Figure 1. Non-measurable criteria categories cited by participants.

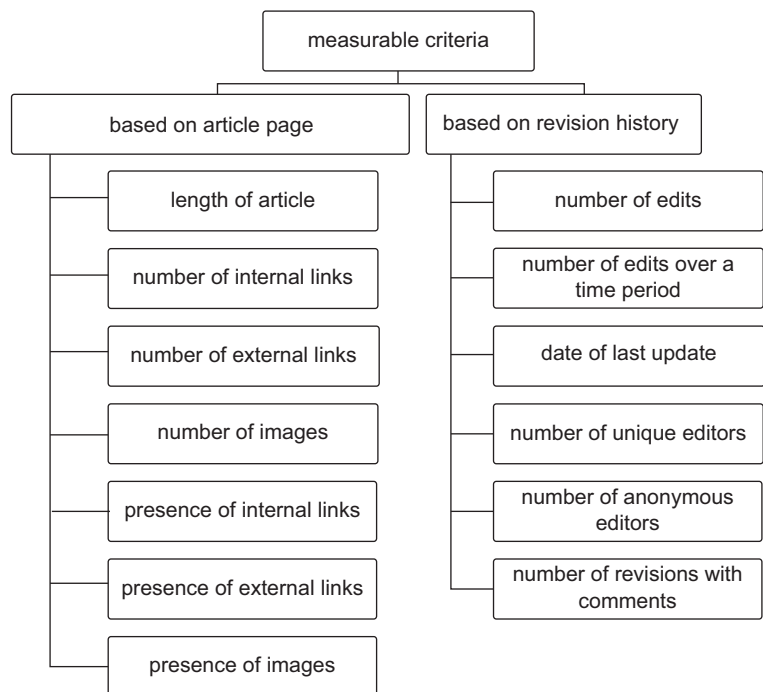


Figure 2. Measurable criteria cited by participants.

The measurable criteria appearing in Figure 2 again are divided into criteria mentioned when viewing the article and the history page. In each of these two subcategories the criteria can be further subdivided into criteria that only note the presence or absence of a quantitative attribute (e.g. presence of internal links) and criteria that relate to the extent or amount of the quantitative attribute (e.g. number of images). The most frequently mentioned quantitative attribute was the length of the article; this criterion was mentioned by 18 participants positively for high-quality articles and by 26 participants negatively (by 28.13% and 40.63% of the participants respectively). An example of a negative comment on an article assessed as low quality is: ‘Because this is not a long article, it is quite possible that it is not comprehensive. I would expect something longer here’; and a positive comment on an article that was perceived as a high-quality article: ‘When so much is written about one article I feel that there is a lot of effort invested in it; when I see so much written, it is great, it raises my expectation for a worthy article’.

Table 1. Distribution of the usage of non-measurable criteria

	Highest quality article		Lowest quality article	
	Number of participants mentioning the criterion	Percentage of participants mentioning the criterion	Number of participants mentioning the criterion	Percentage of participants mentioning the criterion
Based on article page				
coverage	23	35.94%	17	26.56%
structure	12	18.75%	5	7.81%
writing style	5	7.81%	4	6.25%
relevance of internal links	3	4.69%	3	4.69%
quality of external links	3	4.69%	1	1.56%
Based on revision history				
comments on edits	14	21.88%	5	7.81%
attitude towards topic	4	6.25%	5	7.81%
nicknames of editors	2	3.13%	2	3.13%

Table 2. Distribution of the usage of measurable criteria

	Highest quality article		Lowest quality article	
	Number of participants mentioning the criterion	Percentage of participants mentioning the criterion	Number of participants mentioning the criterion	Percentage of participants mentioning the criterion
Based on article page				
length of article	18	28.13%	26	40.63%
number of internal links	4	6.25%	3	4.69%
number of external links	9	14.06%	1	1.56%
number of images	1	1.56%	0	0.00%
presence of internal links	2	3.13%	0	0.00%
presence of external links	13	20.31%	9	14.06%
presence of images	8	12.50%	7	10.94%
Based on revision history				
number of edits	6	9.38%	10	15.63%
number of edits over a time period	9	14.06%	8	12.50%
date of last update	4	6.25%	3	4.69%
number of unique editors	8	12.50%	4	6.25%
number of anonymous editors	1	1.56%	2	3.13%
number of revisions with meaningful comments	2	3.13%	6	9.38%

For high-quality articles, the presence of external links was noted by 13 participants, and nine additional participants commented on the number of external links (22 participants or 34.38% altogether); thus if we add these two subcategories together, the existence of external links becomes the most frequently mentioned attribute of high-quality articles. The lack of external links was noted by 10 participants (15.63%) for articles considered by them to be of low quality. An example of a statement for an article considered to be of high quality is: ‘The fact that there are lots of references here, and that I can go and read them allows me to compare between this article and other sources. Look at the amount of references, whoever wrote this did not just write it, he had sources to rely on’. Here is a negative statement regarding external links on a low-quality article: ‘There are no external links. Just having the options of external links lets me cross check the information’.

For high-quality articles, the presence of images in the articles was viewed as a positive attribute by eight participants (12.50%), and one additional participant noted positively the number of images in the article. The lack of images was mentioned by seven participants (10.94%) for low-quality articles, for example: ‘The content is sufficient but in my

opinion you need pictures because it is more pleasing to the eye. I like pictures a lot'. The presence and number of internal links were mentioned positively by six participants for the high-quality articles, and the lack of internal links was noted by three participants for the low-quality articles.

When viewing the history pages, the number of overall comments was less than for the article pages; however, here we had some interesting findings. The participants had differing interpretations regarding the number of edits and the number of edits over a time period. Some of them (nine participants; 14.06%) viewed a large number of edits as a sign of high quality, while six participants (9.38%) thought that a small number of these are a sign of high quality. The rest of the participants had not commented on this aspect of the article. A quote from a participant who considered the large number of edits as a sign of quality: 'I prefer a page like this, with lots of changes in the article, it shows that there is some sort of development. It is not something static that happened and is now over, but something that can always have additions, something being studied'. And a counterexample, where the participant views a small number of edits as a sign of quality: 'Not too many edits, so it looks like people did not edit the article just because they were bored. When there are too many edits it seems to me that people just entered and perhaps messed it up'.

Similarly, for the article considered by them to be of the lowest quality, eight participants (12.50%) thought that a large number of edits corresponds to low quality (e.g. 'This is least good because it has too many changes, maybe they are not sure of themselves'; 'There are many editors here, everyone joins and says something and adds a bit and ruins a bit, if each person contributed a little, how can you tell who is behind it?') while eight other participants (12.50%) stated that a small number of these attributes corresponds to low quality (e.g. 'There are not many changes here ... no one is interested in it and not enough eyes have looked it over, which means to me that I cannot trust the information').

Some of the participants mentioned the number of edits in general (mentioned for 16 articles when combining comments on high- and low-quality articles), while others noted the weighted version: number of edits over a time period (mentioned for 17 articles when combining comments on high- and low-quality articles). With regards to the number of unique editors, there was more agreement: six participants viewed a large number of editors as a sign of high quality; while only two thought that a small number of editors indicate high-quality. For the low-quality article, all the participants who mentioned this attribute thought that the small number of unique editors indicated low-quality. Overall it seems that a large number of edits and editors are viewed as a sign of quality, but the findings here are not clear cut.

The other attributes on the right side of Figure 2 were mentioned by just a few participants, among these the up-to-dateness of the information seems to be most important: 'As far as I am concerned, articles that are more up-to-date are of higher quality'.

Finally, half of the participants (32 participants) saw the articles with the actual Wikipedia templates. During evaluation, not one of them identified the template indicating the featured article, perhaps because this template is inconspicuous. However, 11 (34.37%) participants noticed the other templates ('rewrite,' 'edit,' 'expand') and only six (18.75%) chose to mention the template as part of their argument. The template most frequently mentioned was 'rewrite', for the lowest-quality article: 'Look at the template; it says that whoever edited this made a comment that there is a possibility of errors. How can you trust this, if there could be mistakes, if you have to study for an exam or hand in a paper?' It should be noted that, of the 21 remaining participants (65.62%) who did not mention the templates as part of their arguments, eight (38.10%) claimed that they had seen the templates but thought they were irrelevant and therefore ignored them: 'I saw it but just skipped over it, because I thought it was some sort of general statement and not relevant to the topic'.

3.2. Community versus user judgement of the quality of Wikipedia articles

Each participant was asked to pick the highest and lowest quality article in his/her opinion out of five Wikipedia articles in the chosen topic. Each participant received one of each: a featured article, an article requiring expansion, an article requiring cleanup (editing), an article needing rewriting, and a regular article without any Wikipedia templates assigned to it. In this section we analyse the correspondence between the participants' choices and the Wikipedia templates. A priori it is reasonable to expect that the users' choice of the highest quality article coincides with the featured article. However, as we can see from Table 3, this was not exactly the case. Although the featured article was chosen most often, this only happened in less than 50% of the cases. In Table 4 we see the distribution of the Wikipedia templates for the article that was chosen to be of the lowest quality by the participants. It is interesting to note that, in three cases, the featured article was chosen as the lowest quality article out of the five articles assessed by these participants. One of the participants who chose the featured article as the lowest quality article complained about the lack of images in the article: 'Provide some images, this is an abstract concept, how can you do without visualization?', and the others found the article too high level and detailed. When considering those who saw the articles with and without Wikipedia templates, we observe that the featured article was chosen as 'best' by considerably more participants when the templates were seen, even though the participants did not acknowledge the influence of these templates. Regarding the 'worst' article, the

Table 3. The distribution of the Wikipedia templates assigned to the highest quality article chosen by the participants, total participants, participants who saw the articles without and with the Wikipedia templates. Percentages in parenthesis

Wikipedia template	Number of times chosen	Number of times chosen – templates removed	Number of times chosen – with templates
Featured	28 (45.75%)	8 (25.00%)	20 (62.50%)
Regular	14 (21.88%)	7 (21.88%)	7 (21.88%)
Expand	8 (12.50%)	7 (21.88%)	1 (3.13%)
Edit	7 (10.94%)	4 (12.50%)	3 (9.38%)
Rewrite	7 (10.94%)	6 (18.75%)	1 (3.13%)

Table 4. The distribution of Wikipedia templates assigned to the lowest quality article chosen by the participants, total participants, participants who saw the articles without and with the Wikipedia templates. Percentages in parenthesis

Wikipedia template	Number of times chosen	Number of times chosen – templates removed	Number of times chosen – with templates
Rewrite	24 (37.50%)	10 (31.25%)	14 (43.75%)
Edit	19 (29.69%)	9 (28.13%)	10 (31.25%)
Expand	11 (17.19%)	7 (21.88%)	4 (12.50%)
Regular	7 (10.94%)	3 (9.38%)	4 (12.50%)
Featured	3 (4.69%)	3 (9.38%)	0 (0.0%)

differences are much less pronounced; however, it should be noted that only those users who did not see the templates chose a featured article as ‘worst’.

4. Discussion

The aim of this study was to explore what attributes users use in order to assess Wikipedia articles. Wikipedia is a collaborative effort, and thus the author or the authors of a Wikipedia article are not known, and therefore the most frequently used traditional criterion, authority, is problematic when assessing Wikipedia articles. Still, we saw that the participants mentioned attributes of Wikipedia articles that are often used for assessing information from the web in general.

The most frequently mentioned evaluation criterion in the present study was the amount of information in an article (length). This criterion was also found as a major criterion by Tillotson [8] and Rieh [4], but not as a major criterion. In the current study we distinguished between the non-measurable criterion of being satisfied with the scope and the content and the measurable criterion relating to the length of the article. In previous works these two aspects were often considered as one (e.g. [4,8]). Participants’ statements citing the length of an article as an indicator of quality were analysed. The analysis revealed that their view of length as a quality indicator was not entirely detached from the content. In other words, seeing the connection between length and quality is based on the fact that quantity can indicate comprehensiveness, so that it is more probable that a long text will be more detailed, exhaustive, and comprehensive than a short text. This perception can be viewed as an application of the ELM model [13] – if we cannot assess the degree to which the content is comprehensive and exhaustive, we use this rule of thumb (the length of the article) to deduce something about its quality.

Blumenstock [31] suggested the length of Wikipedia articles as a proxy for quality, based on the observation that featured articles are longer than the regular ones. Here, however a word of caution is necessary. Featured articles undergo a careful reviewing process by the community, where one of the requirements for becoming a featured article is its comprehensiveness, without going into excessive detail [43]. However, excessive detail and the length of an ideal Wikipedia article are somewhat subjective [44]. The review process also increases the number of edits, because the reviewers enhance the article, so that it fulfils all the requirements: well-written, comprehensive, well-researched, neutral and stable. Thus it is not surprising that featured articles are longer, have more edits and editors on average compared with regular articles.

In addition to the amount of information, the presence and/or quantity of external links was also often used to assess the quality of web sites in general. Links and references were mentioned in [9, 10], for example. It is notable that in the present study, as in those of Tillotson [8] and Freeman and Spyridakis [10], in most cases when this attribute was mentioned, the participants did not address the quality of the external links. Thus, what was important to them was neither the type of source that the link related to, nor its quality – they only cared about the presence of links. The existence of

external links was viewed as slightly more important than the number of external links. It is well known [4] that the presence of sources that allow verification of written content is one of the key factors in assessing reliability. We assume that an article that contains external links is perceived as being of higher quality because the references invite readers to check that the content is accurate, thus making users feel more secure. In addition, external links are viewed as sources for further information about the subject matter discussed in the article [10]. It is possible that the users' basic assumption was that external links in Wikipedia environment are of high quality. Their presence spared them the effort of using a search engine to locate additional sources and to assess them. This explains why the users commented not only on the presence or absence of external links but on the number of links as well. Encyclopaedia articles are often viewed as starting points for learning about a topic, and users expected to be guided to sources for further exploration.

Regarding images, it seems that the number of images in the article is not a measure of its quality, and what is important is the fact that they are there. The arguments for selecting this attribute show that the saying 'a picture is worth a thousand words' is not a mere cliché. Rather, as revealed by Rieh [4], the visual aids that accompany the written information have a certain influence on the perception of the quality of the text, as they assist in understanding the content.

In addition to examining the content, the participants were asked to evaluate the article's quality by examining its development as revealed by the history page. The participants especially noted those attributes that were related to the number of edits the article had undergone. The participants used two approaches which differed in the way they reached a conclusion regarding the attribute, but the frequency of use was similar. The first approach was based on a general assessment of the number of changes that had appeared on the screen, and the other weighted this number against the age of the article (number of edits over a period of time).

Rather interestingly, the number of edits received two contradictory interpretations by the participants. Some users were guided by the principle that many edits enrich the content and enhance it. Others felt that the opposite was true, and that the many edits diminish the article's quality, and therefore fewer edits are more desirable. Another reason for preferring a small number of edits and editors was the assumption that the small number of edits point to the high quality of the article, as the community did not find it necessary to improve it. These findings support the claim that quality is a subjective concept which depends on the user's unique viewpoint. It should be noted that most participants had never seen the history pages, and it is possible that their attitude towards the number of edits and editors was not yet formed and could change as they learn more about the way Wikipedia works. In terms of the number of editors, the users showed preference for a larger number of editors, especially when they saw that the involvement of the editors was ongoing. Some participants viewed the large number of editors as a sign of quality, an application of the 'wisdom of crowds' principle.

A large number of Wikipedia evaluation studies based their assessments on the number of edits, number of editors and the number of edits over a time period (e.g. [18, 23, 29]). In these studies it is usually assumed that more is the better, although Kittur and Kraut [26] found that increasing the number of editors improves quality only when the work between them is coordinated, either explicitly or implicitly.

Most studies compared characteristics of featured articles with those of regular Wikipedia articles, or sometimes with articles in other graded categories assigned by the Wikipedia community (e.g. [19, 23, 25, 26, 29]). Our study is based on actual user assessments; we saw that the users considered the featured article as the highest quality article in only slightly more than 45% of the cases. In addition the participants' interpretation of the number of edits did not coincide with the interpretation in the studies, suggesting quality measures based on the number of edit, i.e. for some of the users, more is not necessarily better. This finding warrants a re-examination of the previously suggested quality measures.

In the present study only a few participants mentioned currentness when they evaluated the quality of articles, perhaps because they take it for granted that Wikipedia articles are up-to-date.

5. Summary and conclusion

In the current study non-expert users were asked to judge Wikipedia articles. Based on their justifications for choosing the best and the worst article, evaluation criteria were defined. The examination of the way users cope with the task of assessing community-generated content revealed that they adapt to the environment they examine. They are capable of finding ways to evaluate the reliability and quality of the content even when traditional evaluation criteria are not valid or are difficult to apply.

The setting of this study was not naturalistic, thus it is not certain that users actually apply these criteria when evaluating Wikipedia articles. On the other hand, we are not aware of other studies that identified Wikipedia article evaluation criteria from actual user judgements, and without providing a list of possible criteria (like in questionnaires). We view this as an important contribution of the current work, especially since our findings showed that criteria suggested in previous work (like number of edits) are not always viewed by the users the same way as they are viewed by researchers. The perceptions of our users regarding quality did not always coincide with the perceptions of Wikipedia editors, since

in fewer than half of the cases the featured article was chosen as best. This finding warrants further exploration. Previous studies often relied implicitly on the high quality of featured articles. An additional novel finding is that users have differing opinions as to the meaning of a large number of edits. In line with the wisdom of crowds, more edits are a sign of quality, while on the other hand few edits may mean that the information is already of high quality and there is no need to improve it.

This study has limitations, since this was an exploratory study. The participants were all university students, not randomly selected from the general user population of the internet or Wikipedia. Information science students, the large majority of the participants in the study, probably are more information literate than the general student population. Thus one has to be very careful about generalizing the results even to Israeli students. Wikipedia articles came from the Hebrew language Wikipedia, and the results might not be generalizable to Wikipedia in other languages. In addition the perceptions of the Israeli users might not reflect the perceptions of other populations.

It is important to note that in the present study users considered only the content page and history pages. Dondio et al. [30] and Wilkinson and Huberman [25] showed that data on discussion page (talk pages) contribute to indentifying the quality of articles in this environment, and discussion pages were also mentioned by Nofrina et al. [35]. Therefore, it is recommended to include the discussion page in future information quality evaluation studies.

Acknowledgement

The study reported here is based on the doctoral research of the first author conducted in the Department of Information Science, Bar-Ilan University, under the supervision of Professor Shifra Baruchson-Arbib, and was supported by the Bar-Ilan's President Scholarships Program.

References

- [1] American Library Association. *Presidential Committee on Information Literacy: final report*. Available at: <http://www.ala.org/ala/mgrps/divs/acrl/publications/whitepapers/presidential.cfm> (accessed 27 February 2011)
- [2] Janes JW, Rosenfeld LB. Networked information retrieval and organization: Issues and questions. *Journal of the American Society for Information Science*, 1996; 47(9): 711–715.
- [3] Metzger MJ. 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* 2007; 58(13): 2078–2091.
- [4] Rieh SY. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology* 2002; 53(2): 145–161.
- [5] Arazy O, Kopak R. On the measurability of information quality. *Journal of the American Society for Information Science and Technology* 2011; 62(1): 89–99.
- [6] Hilligoss B, Rieh SO. Defining a unifying framework for credibility assessment: construct, heuristics and interaction in context, *Information Processing and Management* 2008; 44: 1476–1484.
- [7] Fink-Shamit N, Bar-Ilan J. Information quality assessment on the Web – an expression of behaviour. *Information Research* 2008; 13(4): paper 357. <http://InformationR.net/ir/13-4/paper357.html> (accessed 11 March 2011).
- [8] Tillotson J. Web site evaluation: a survey of undergraduates. *Online Information Review* 2002; 26(6): 392–403.
- [9] Eysenbach G, Köhler C. How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 2002; 324: 573–577.
- [10] Freeman KS, Spyridakis JH. An examination of factors that affect the credibility of online health information. *Technical Communication* 2004; 51(2): 239–263.
- [11] Fogg BJ, Soohoo C, Danielson DR, Marable L, Stanford J, Tauber ER. *How do users evaluate the credibility of web sites? A study with over 2,500 participants*. Proceedings of the 2003 conference on Designing for user experiences, San Francisco, CA; 1–15.
- [12] Rieh SY, Danielson DR. Credibility: a multidisciplinary framework. In: Cronin B, editor. *Annual Review of Information Science and Technology*, 41. Medford, NJ: Information Today, 2007: 307–364. Online Referencing http://www.si.umich.edu/rieh/papers/rieh_ARIST2007.pdf (accessed 11 March 2011).
- [13] Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology* 1986; 19: 123–205.
- [14] Alexa. *Top sites*. Available at: <http://www.alexa.com/topsites> (accessed 8 June 2011).
- [15] Surowiecki J. *The wisdom of crowds*. New York: Anchor Books, 2005.
- [16] Giles J. Internet encyclopaedias go head to head. *Nature* 2005; 438(7070): 900–901.
- [17] Rector LH. Comparison of wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review* 2008; 36(1): 7–22.
- [18] Lih A. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. *Proceedings of the 5th international symposium on online journalism*, Austin, TX, 2004.

- [19] Zeng H, Alhossaini MA, Ding L, Fikes R, McGuinness DL. Computing trust from revision history. In: *Proceedings of the international conference on privacy, security and trust*. Ontario: Markham, 2006.
- [20] *Wikipedia: Featured articles*. Available at: http://en.wikipedia.org/wiki/Wikipedia:Featured_articles (accessed 8 June 2011).
- [21] Adler BT, de Alfaro L. A content-driven reputation system for the Wikipedia. In: *Proceedings of WWW'07*, 2007: 261–270.
- [22] Adler BT, Chatterjee K, de Alfaro L, Faella M, Pye I, Raman V. Assigning trust to Wikipedia content. In: *Proceedings of the 4th international symposium on wikis*. Porto: ACM, 2008.
- [23] Hu M, Lim EP, Sun A, Lauw HD, Vuong BQ. Measuring article quality in Wikipedia: models and evaluation. In: *Proceedings of CIKM07*. Lisbon: ACM, 2007.
- [24] *Wikipedia: Version 1.0 Editorial Team/Assessment*. Available at: http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment (accessed 8 June 2011).
- [25] Wilkinson D, Huberman B. Cooperation and quality in the Wikipedia. In: *Proceedings of WikiSym*, Montreal, 2007.
- [26] Kittur A, Kraut RE. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: Begole B, McDonald DW (eds), *Proceedings of the 2008 ACM conference on computer supported cooperative work*. San Diego, CA: ACM, 2008.
- [27] Wöhner T, Peters R. Assessing the quality of Wikipedia articles with lifecycle based metrics. In: *Proceedings of WikiSym'09*, Orlando, FL, 2009.
- [28] Stein K, Hess C. Does it matter who contributes? – a study on featured articles in the German Wikipedia. In: *Proceedings of HT'07*. Manchester, 2007: pp. 171–174.
- [29] Stvilia B, Twidale MB, Smith LC, Gasser L. Assessing information quality of a community-based encyclopedia. In: Naumann F, Gertz M, Mednick S (eds), *Proceedings of the international conference on information quality – ICIQ*, Cambridge, MA, 2005: 442–454.
- [30] Dondio P, Barrett S, Weber S, Seigneur JM. *Extracting trust from domain analysis: a case study on the Wikipedia project*. Lecture Notes in Computer Science 4158. Berlin: Springer-Verlag, 2006. Available at: <http://www.springerlink.com/content/ump230u335h4nh97/> (accessed 27 February 2011).
- [31] Blumenstock JE. Size matters: word count as a measure of quality on Wikipedia. In: *Proceedings of the 17th international conference on World Wide Web*, Beijing. ACM, 2008.
- [32] Kittur A, Chi ED, Suh B. Crowdsourcing user studies with Mechanical Turk. In: *Proceedings of CHI 2008*, Florence, 2008.
- [33] Pirolli P, Wollny E, Suh B. So you know you're getting the best possible information: a tool that increases Wikipedia credibility. In: *Proceedings of CHI'09*, Boston, MA, 2009: 1505–1508.
- [34] Chesney T. An empirical examination of Wikipedia's credibility. *First Monday* 2006: 11(11) Available at: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1413/1331> (accessed 27 February 2011).
- [35] Nofrina H, Viswanathan V, Poorisat T, Detenber BH, Chen P. Why some wikis are more credible than others: structural attributes of collaborative websites as credibility cues. *OBS Journal*, 2009; 9. Available at: <http://obs.obercom.pt/index.php/obs/article/view/261> (accessed 7 June 2011).
- [36] Metzger MJ, Flanagin AJ, Medders R. Social and heuristic approaches to credibility evaluation online. *Journal of Communication* 2010; 60(3): 413–439.
- [37] *Wikipedia contributors. Hebrew Wikipedia*. Available at: http://en.wikipedia.org/wiki/Hebrew_Wikipedia (accessed 7 June 2011).
- [38] *Wikipedia: Statistical data* (in Hebrew). Available at: http://he.wikipedia.org/wiki/%D7%95%D7%99%D7%A7%D7%99%D7%A4%D7%93%D7%99%D7%94:%D7%A0%D7%AA%D7%95%D7%A0%D7%99%D7%9D_%D7%A1%D7%98%D7%98%D7%99%D7%A1%D7%98%D7%99%D7%99%D7%9D/%D7%97%D7%95%D7%93%D7%A9%D7%99%D7%AA (accessed 8 June 2011).
- [39] Parag N. *Israelis like Wikipedia* (in Hebrew). Available at: <http://www.mako.co.il/digital-magazine/Article-0035fbac8a0121004.htm> (accessed 8 June 2011).
- [40] Ranie L, Tancer B. *Wikipedia, when in doubt multitudes seek it out*. Pew Internet and American Life Project, 2007. Available at: <http://pewresearch.org/pubs/460/wikipedia> (accessed 8 June 2011).
- [41] Patton MQ. *Qualitative research & evaluation methods* (3rd edn). Thousand Oaks, CA: Sage, 2002.
- [42] Ericsson KA, Simon HA. *Protocol analysis: verbal reports as data* (revised edn). Cambridge, MA: The MIT Press, 1993.
- [43] Wikipedia Contributors. *Wikipedia: Featured article criteria*; 2010. Available at: http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria (accessed 27 February 2011).
- [44] *Wikipedia contributors. Wikipedia: summary style*, 2010. Available at: http://en.wikipedia.org/wiki/Wikipedia:Summary_style (accessed 27 February 2011).