

So You Know You're Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility

Peter Pirolli

Palo Alto Research Center
3333 Coyote Hill Rd,
Palo Alto, CA 94304, USA
pirolli@parc.com

Evelin Wollny

Knowledge Media Research Center
Konrad-Adenauer-Strasse 40,
Tuebingen, D--72072, Germany
e.wollny@iwm-kmrc.de

Bongwon Suh

Palo Alto Research Center
3333 Coyote Hill Rd,
Palo Alto, CA 94304, USA
suh@parc.com

ABSTRACT

An experiment was conducted to study how credibility judgments about Wikipedia are affected by providing users with an interactive visualization (WikiDashboard) of article and author editing history. Overall, users who self-reported higher use of Internet information and higher rates of Wikipedia usage tended to produce lower credibility judgments about Wikipedia articles and authors. However, use of WikiDashboard significantly increased article and author credibility judgments, with effect sizes larger than any other measured effects of background media usage and attitudes on Wikipedia credibility. The results suggest that increased exposure to the editing/authoring histories of Wikipedia increases credibility judgments.

Author Keywords

Wikipedia, credibility, WikiDashboard

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Wikipedia...is the best thing ever. Anyone in the world can write anything they want about any subject. So you know you are getting the best possible information. – An ironic claim by Michael Scott (played by Steve Carrell), NBC's "The Office."

Although Wikipedia is popular, it has, in its short history, been the target of skepticism (or at least diminished expectations) about the quality of its content and contributors. The conventional stereotype many people have had about great reference works such as encyclopedias was that they were the products of highly reputable experts writing about the things they had devoted their lives to understanding. One reason why media headlines highlighted Nature's [5] comparison of articles in

Wikipedia and Encyclopedia Britannica was that the study found little substantial difference in accuracy between the two sources, and this ran counter to the conventional wisdom of what makes for superior quality reference information. Additional studies [2] indicate that many Wikipedia articles are judged as credible by experts, yet there remains wide-spread skepticism. If high credibility information is the aim of Wikipedia (and wikis more generally), then it is important that users efficiently perceive the credibility of the contributed content.

Most of the more popular Wikipedia articles have a complex history of edits and revisions by many editors. Some historical data are available through tabs at the top of Wikipedia article pages (e.g., "history" and "discussion"), which can in turn lead to data about users (on "user" pages). The "standard" wiki interfaces make such historical data available, but many navigation steps are required to gain a general sense of the article history, and the history of its editors. Users' credibility assessments could be enhanced by better support for making sense of the editing and article histories.

WikiScanner¹, WikiRage², and History Flow³, are all systems that attempt to provide user interfaces that improve on the standard wiki interface and provide greater transparency about the history of Wikipedia articles. The WikiDashboard tool [10] is an interactive visualization that is expected to improve the ease with which users can access and make sense of data about articles and editors. Here we investigate how the use of WikiDashboard affects credibility judgments about Wikipedia articles and authors.

WIKIDASHBOARD

WikiDashboard [10] was developed to provide users of wikis with improved access to cues about the editing history of Wiki articles and their editors. WikiDashboard provides visualizations embedded within Wikipedia pages, along with the live content from Wikipedia. The prototype can be used just as if users are on the Wikipedia site itself. The prototype provides two types of dashboards:

¹ wikiscanner.virgil.gr/

² www.wikirage.com/

³ www.research.ibm.com/visual/projects/history_flow/

1. *Article Dashboards* (Figure 1), which are embedded within articles and which display aggregate edit activity graphs representing the weekly edit trend of the article, followed by a list of the top active editors for that page. The active users of the article are ordered by the number of edits they have made. A weekly edit activity graph for each editor on the right side of the dashboard enables users to investigate when the edits by that editor were made.
2. *User Dashboards* (Figure 2), which are embedded within user pages, which display information relating to a user. In WikiDashboard, each user page has a User Dashboard embedded, displaying the article contribution and editing patterns of that user. The top summary graph shows the editor's weekly edit activity, which allows users to easily examine the editor's overall edit patterns. The summary graph is followed by the list of Wikipedia pages on which the editor has made edits. The list is ordered by the volume of contribution and includes the corresponding article-editor activity graphs on the right side.

Article titles, user names, and statistics in both dashboards are clickable links, allowing users to browse through them for further exploration.

CREDIBILITY

Modern research on credibility dates to the work of Hovland [6] and can be found in fields including psychology, communication, marketing, management, information science, and human-computer interaction. With the rise of online media, such credibility research has included the Internet [3], the Web [4], and Wikipedia [2]. Rieh and Danielson [8] provide a comprehensive review across disciplines.

Historically [8] two core components of the hypothetical construct of credibility have been (1) expertise and (2)

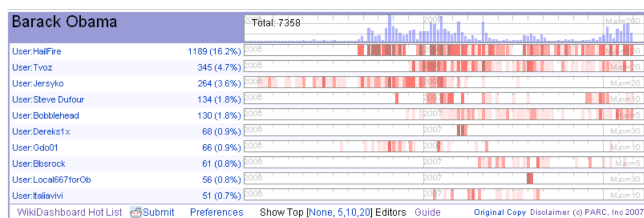


Figure 2 Article Dashboard. The top summary graph shows the weekly edit trend of this page. Below the summary graph, there is a list of active editors and their activities on the article.

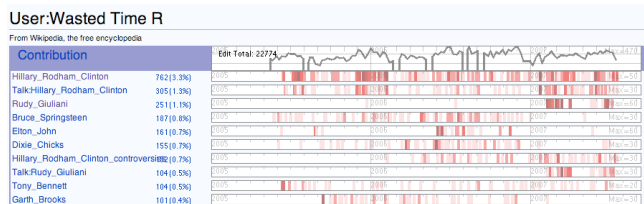


Figure 2 User Dashboard. The dashboard displays weekly edit trend of an editor as well as the list of articles that the editor made revisions on.

trustworthiness. Empirical studies of credibility judgments typically attempt to assess participants' credibility judgments about the information itself (e.g., an article), the sources (e.g., authors), or the media (e.g., Wikipedia as a whole). We build upon the work of Chesney [2] who developed credibility questionnaires as instruments for measuring judgments of article credibility and author credibility. We also included a battery of questions to assess media usage patterns and attitudes to determine if these pre-experimental biases also influence Wikipedia credibility judgments.

RELATED WORK

Kittur et al. [7] showed that non-interactive (static) visualizations similar to Figures 1 and 2 could impact trust judgments. That work used two artificial representations designed to convey what the researchers hypothesized would be "high trust" or "low trust" cues (primarily whether the edit rates were increasing or decreasing). These two visual representations were contrasted with a no-representation baseline condition by associating the representations with articles. As hypothesized, the "high trust" visualization boosted "trust" assessments relative to the baseline no-visualization condition, whereas the "low trust" visualization decreased "trust" assessments.

The current experiment extends that result in several ways. First, in the current experiment the fully interactive WikiDashboard provided representations of the actual article and editor histories (not an artificial set of cues). Second, WikiDashboard was contrasted with a baseline condition in which users were encouraged to interact with the same editing history data available via the standard wiki interface. Third, questions culled from prior credibility research were used to assess both article and author credibility, which is a broader construct than just trust assessment. Finally, the current experiment related the credibility assessments to users' background knowledge and beliefs.

METHOD

Design

Across two groups, we contrasted the use of the WikiDashboard + Wikipedia against a baseline control condition involving just the Wikipedia interface. We also tested (within groups) two kinds of articles (1) Skeptical articles that had been independently identified as being of skeptical credibility and (2) non-Skeptical articles of equivalent lengths were not on the skeptical list. This results in a 2 (WikiDashboard vs No-WikiDashboard) × 2 (Skeptical vs Non-skeptical) mixed factorial design.

If the Skeptical vs Non-skeptical manipulation produces differences on our credibility judgment tests then we have some confirmation for the test validity. In addition, the interaction (or lack of interaction) with the Wikidashboard vs No-Wikidashboard factor allows us to determine if Wikidashboard polarizes further (or not) the differences between Skeptical and Non-skeptical credibility judgments.

Credibility	Article Type	
	Non-skeptical	Skeptical
Article	4.29	3.40
Author	3.81	3.29

Table 1. Mean credibility scores by article type (max = 5).

Credibility	Interface Type	
	No-WikiDashboard	WikiDashboard
Article	3.74	3.95
Author	3.48	3.61

Table 2. Mean credibility scores by interface types (max = 5)

Participants

Twenty-four participants were recruited from the Palo Alto Research Center (as volunteers) and craigslist.com (for \$25 compensation). Participants were randomly assigned to the WikiDashboard or No-WikiDashboard groups.

Articles

A total of six articles were presented to each participant. Three Skeptical articles were chosen from a list on the *WikiProject Rational Skepticism* page on Wikipedia: *universal intelligence*, *grapefruit seed extract*, and *extrasensory perception*. Three Non-skeptical articles of similar lengths were selected from the *WikiProject Germany* page of Wikipedia: *Leipzig*, *Dresden Frauenkirche*, and *Chancellor of Germany*.

Credibility Questionnaire

Article and author credibility was assessed with a set of 12 questions derived from prior credibility studies [2, 3].

An *article credibility* test was composed of six article credibility questions that asked users to rate the following on a 5-point Likert scale ranging from 1="strongly disagree" to 5="strongly agree." Each of these questions began, "I perceived the information to be..." and ended with (1) "believable", (2) "accurate", (3) "trustworthy", (4) "unbiased", (5) "complete", (6) "understandable."

An *author credibility* test was composed of six author credibility questions that asked users to rates on a similar 5-point Likert scale, "I perceived the writer(s) to ...": (1) "be credible", (2) "have high integrity", (3), "have positive reputation", (4) "be trustworthy", (5) "offer information of superior quality", and (6), "be prestigious."

Media Usage and Attitudes Questionnaire

Media usage patterns were assessed using 5-point Likert scale items. Usage rates of computer, Internet, email Wikipedia, search, and word processors, was queried using 6-point Likert scale items. Attitudes towards information sources, including textbooks, Internet, newspapers, television, radio, and magazines, (with "other" as a residual category), were assessed using 5-point Likert items.

Procedure

Participants received a brief introduction to the credibility judgment task and to interaction with WikiDashboard or

Wikipedia history, discussion, and user tabs. This was done using a Wikipedia article for San Francisco. Users also practiced answering all of the article and author credibility questions.

During the main phase of the experiment, participants went through six trials. On each trial they read an article, interacted with WikiDashboard or the standard wiki interface, and then answered the 12 credibility questions. Presentation order of the six articles was arranged in a Latin Square counterbalancing over blocks of six subjects

RESULTS

The credibility tests both showed high internal consistency: author credibility test Cronbach's $\alpha = 0.79$ and article credibility test Cronbach's $\alpha = 0.77$. For the remainder of our analyses credibility scores were computed as the means of the item scores making up each credibility test. We employed a technique recommended by Snell [9] to transform the raw Likert responses to better fit the assumptions required for our statistical analyses below.

Table 1 presents the mean credibility scores for the two different types of articles (Non-skeptical vs Skeptical). The credibility of Non-Skeptical articles was significantly greater than the Skeptical articles, $F(1, 22) = 31.64, p < 0.00005, MSE = 9.99$ and Non-Skeptical authors were judged more credible than Skeptical authors, $F(1, 22) = 12.36, p < 0.005, MSE = 11.81$. Article type had no interactions with the No-WikiDashboard vs WikiDashboard factor on article or author credibility. Given this lack of interaction, the remaining analyses on effects of the WikiDashboard pooled together credibility scores on the different types of articles. We discuss this lack of interaction further in the conclusion section.

Table 2 presents the credibility scores for the interfaces. The absolute differences appear small, but the WikiDashboard credibilities are greater, as we shall discuss. Preliminary exploratory factor and correlation analyses showed relations between media usage patterns and the credibility tests, we performed an analysis of covariance to test the effects of WikiDashboard that included background and media usage responses as covariates.

Tables 3 and 4 present results from analyses of covariance of the mean article and author credibility scores. The linear models for each type of credibility scores included the interface type (WikiDashboard or No-WikiDashboard) and all of the Media Questionnaire responses (no interactions were included). For brevity, only the marginally significant and statistically significant effects are presented. The use of WikiDashboard produced statistically significant increases in article credibility and a borderline significant ($p = .05$) increase in author credibility. Interestingly, users who report higher use of the Internet as an information source have statistically lower credibility scores for Wikipedia author and article credibility. In a similar vein, participants who reported higher use of Wikipedia also tended to produce lower credibility scores, however, of those people

who self-report as regular Wikipedia users, those with higher rates of reading Wikipedia articles produced higher credibility assessments. Of all the effects listed in Tables 3 and 4, the presence of WikiDashboard produced the largest absolute effect size.

Coefficient	Estimate	Std. Error	t	p
WikiDashboard	0.95	0.30	3.15	0.016
Info from Internet	-0.51	0.19	2.74	0.029
Use Wikipedia	-0.53	0.20	2.62	0.034
Read Wikipedia	0.65	0.20	3.18	0.016

Table 3. Article credibility: Coefficient estimates and statistics for statistically significant and marginally significant effects.

Coefficient	Estimate	Std. Error	t	p
WikiDashboard	0.91	0.39	2.333	0.0524
Info from Internet	-0.75	0.24	3.140	0.0164
Use Wikipedia	-0.57	0.26	2.191	0.0646
Read Wikipedia	0.70	0.26	2.663	0.0323

Table 4. Author credibility: Coefficient estimates and statistics for statistically significant and marginally significant effects.

CONCLUSIONS

WikiDashboard use equally increased the credibility judgments about articles designated as Skeptical or non-Skeptical. It seems likely that the lower credibility scores for the Skeptical articles reflect pre-experimental biases about the subject matter (ESP, universal intelligence, grape seed extract) and this baseline credibility was improved by WikiDashboard exposure to the article and author histories.

Wikidashboard did not increase the polarization of judgments about Skeptical and non-Skeptical articles and authors, and one might question whether such as a tool is useful if it is increasing the credibility of an article whether it is a topic that can attract skepticism (e.g., ESP) or not (e.g., German culture). One interpretation might be that Wikidashboard actually *decreases* critical reading. However, both groups were motivated to read critically and both groups had training on using the standard edit history information. The Wikidashboard group had additional UI representations of the crucial information. We suspect that the increase in credibility judgments is some variation of effects predicted for information asymmetry in “lemon markets” in economics [1]: When presented with a high-quality good about which they have little information, consumers will tend to initially expect the quality to be lower. Providing them with more information signals about a high quality good will raise their expectations of quality. Wikidashboard provides all the information available in the “baseline” case, plus the additional information in the visualization.

Our Media usage and Attitudes questionnaire suggested that people who frequently use the Internet and Wikipedia have a tendency to have lower credibility assessments of Wikipedia articles and authors. In some sense, “familiarity breeds contempt” at least with respect to predispositions to the overall general credibility of Wikipedia articles and

authors, but this appears to be mitigated by providing users with more transparent visualization of the particular authoring history of an article. The boost to credibility incurred by WikiDashboard use is greater than the negative impact of these background biases.

Prevalent skepticism among the public remains about the credibility of Wikipedia articles despite studies [2, 5] that suggest that much of the information is of high quality and reliability. WikiDashboard was designed to provide improved access to article and author editing histories and we found that use of WikiDashboard increased credibility judgments about Wikipedia articles and authors.

ACKNOWLEDGMENTS

This work was supported in part by Office of Naval Research Contract No. N00014-08-C-0029 to Peter Pirolli. Thanks to Gregorio Convertino for comments.

REFERENCES

1. Akerlof, G.A. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84 1970, 488-500.
2. Chesney, T. An empirical examination of Wikipedia's credibility. *First Monday*, 11 (11) 2006.
3. Flanagin, A.J. and Metzger, M.J. Perceptions of Internet information credibility. *Journalism and Mass Communication Quarterly*, 77 (3) 2000, 515-540.
4. Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J. and Tauber, E.R. 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*, ACM, San Francisco, California, 2003.
5. Giles, J. Internet encyclopaedias go head to head. *Nature*, 438 2005, 900-901.
6. Hovland, C.I., Janis, I.L. and Kelley, H.H. *Communication and persuasion*, New Haven, CT: Yale University Press, 1953.
7. Kittur, A., Suh, B. and Chi, E.H. Can you ever trust a Wiki/ Impacting perceived trustworthiness in Wikipedia 2008 *ACM Conference on Computer Supported Cooperative Work, CSCW 2008*, ACM Press, San Diego, CA, 2008.
8. Rieh, S.Y. and Danielson, D.R. Credibility: A multidisciplinary framework. in *Annual Review of Information Science and Technology*, Information Today, Medford, NJ, 2007, 307-364.
9. Snell, E.J. A scaling procedure for ordered categorical data. *Biometrics*, 20 (3) 1964, 592-607.
10. Suh, B., E.H.Chi, Kittur, A. and Pendleton, B.A. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM, Florence, Italy, 2008

