Your Process Is Showing: Controversy Management and Perceived Quality in Wikipedia

W. Ben Towne,[1] Aniket Kittur,[2] Peter Kinnaird,[2] & James Herbsleb[1]
[1]Institute for Software Research, [2]Human-Computer Interaction Institute; School of Computer Science
Center for the Future of Work; Heinz College
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213 USA
{wbt, nkittur, kinnaird, jdh}@cs.cmu.edu

ABSTRACT

Large-scale collaboration systems often separate their content from the deliberation around how that content was produced. Surfacing this deliberation may engender trust in the content generation process if the deliberation process appears fair, well-reasoned, and thorough. Alternatively, it could encourage doubts about content quality, especially if the process appears messy or biased. In this paper we report the results of an experiment where we found that surfacing deliberation generally led to decreases in perceptions of quality for the article under consideration, especially – but not only – if the discussion revealed conflict. The effect size depends on the type of editors' interactions. Finally, this decrease in actual article quality rating was accompanied by self-reported improved perceptions of the article and Wikipedia overall.

Author Keywords
User-generated content; discussion; dispute resolution; trust; perceptions of quality

ACM Classification Keywords
H.5.3 Web-based interaction; Asynchronous interaction; Collaborative computing; Computer-supported cooperative work.

General Terms
Human Factors; Experimentation; Wikipedia; Commons-based Peer Production; Deliberation

INTRODUCTION

Technological advances in recent years, such as wikis, have enabled large-scale systems for aggregating knowledge and information from a very large number of participants, who bring a broad range of viewpoints, information, and expertise. In some systems, like Wikipedia, a significant amount of coordination communication is employed to effectively combine these contributions [11].

However, many readers of Wikipedia are unaware that this work that has gone into the creation of an article, even though these efforts represent a potential source for readers to understand and evaluate the trustworthiness of an article. For example, imagine a reader who encounters a controversial topic in Wikipedia. In one case, the reader sees only the article and must evaluate the likely bias and validity of the topic on its own. Alternatively, the reader also sees that substantial and considered discussion has taken place among the editors of the topic on how to sensitively and appropriately present it. In the latter case, the reader has additional information to judge the article quality. If the discussion seems measured and fair, the reader's evaluation of the article may improve. If, on the other hand, the reader sees unchecked biases or personal attacks among the contributors, the reader may be less likely to trust the content than if it were encountered in isolation. Readers may simply become overwhelmed by the amount of information needed to understand the article creation process. More fundamentally, increasing the visibility of the uncertain and messy process by which articles are created may undermine readers' perceptions of trust — even if that process leads to a preferred outcome [21:1–14, 111–138]. To quote John G. Saxe, "Laws, like sausages, cease to inspire respect in proportion as we know how they are made" [32]. The same may be true of user-generated content.

To examine this question, we conducted an experiment in which we surfaced various types of discussions along with content, and measured readers' perceptions of the quality of the article excerpt being discussed. We found that when discussion was provided alongside content, the quality ratings for the content were significantly lower than when no discussion was displayed, supporting the "sausage" hypothesis. When discussion involving conflict was displayed, article quality ratings were even lower. However, if the editors involved in the conflict resolved it through a positive collaboration approach, the negative effects of conflict disappeared. Participants were not generally aware of the rating-lowering effect of the discussion, and generally reported that reading the discussion raised their perceptions of both the article's quality and Wikipedia in general.

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.
Copyright 2013 ACM 978-1-4503-1331-5/13/02...$15.00.

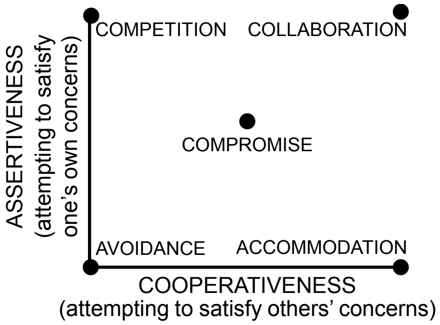

Figure 1: Two dimensional taxonomy of conflict handling modes (adapted from [24])

CONFLICT RESOLUTION STRATEGIES

Conflict has been studied from a variety of different perspectives, mostly in the context of small groups. Thomas and Kilman [24] described a framework in 1976 characterizing approaches to managing conflict, by the degree to which individuals attempt to satisfy their own concerns ("assertiveness") vs. others' concerns ("cooperativeness"). They find the five distinct conflict management strategies shown in Figure 1.

This framework has been adopted in the literature and validated several times [24:269]. It is supported by a number of other studies that independently attempt to build typologies of conflict resolution strategies and identify aspects of conflict management that are important to outcomes, such as Klein & Lu [12]'s early analysis of approaches to solving task conflicts in a cooperative interdisciplinary design team. Klein & Lu's primary conclusions are that "conflict resolution plays a central role in cooperative design…and knowledge acquisition in cooperative design presents special challenges and requires special techniques." The authors believe that new practices with parallel interaction amongst diverse concerns cannot happen without effective conflict resolution [12:169].

In this study, we examine how revealing details of the joint production process affect perceived quality of outcomes. Task conflicts have a wide range of potential effects, both positive and negative. Therefore, we explored only task conflicts, as opposed to relationship conflicts that are generally only seen to have negative effects (see e.g. [6]).

WIKIPEDIA TALK PAGES

Wikipedia is an open and free encyclopedia that anyone can edit. The encyclopedic content and editing process are well studied as an example of a large-scale open collaborative system. A "discussion" tab in the upper left corner of most article pages links to a "Talk" page where editors discuss changes to the article [30].

Viégas et al. [26] confirm Kittur et al's finding [11] that Talk and Project pages are the fastest growing parts of Wikipedia, especially with more heavily edited articles. They coded subsets of 25 article Talk pages (excluding archives) to find out what people are doing there, producing a typology that guided our choice of discussion types. They chose controversial and non-controversial topics in areas from hard science to pop culture, especially including cases with difficult coordination issues. Postings were coded for 11 binary dimensions, which were analyzed for frequency.

The most common kind of posting found in this study is "requests for coordination," with contributors asking for help and explaining why they think specific changes should be made. Over half of Talk page contributions fit this category, including 97% of the discussions coded from the Yasser Arafat page.

The next most common use was a "request for information," found in just over 10% of posts. Writers of these posts hope to tap into the knowledge of an "approachable community of experts" [26:8] on a specific topic, without necessarily having intention to edit the article. The overwhelming majority of these requests were answered, with information or links that might answer the question.

The third most common type of posting was coded as "off-topic remarks," generally users sharing trivia or personal experiences related to the article topic. The fourth most common type of posting, with 7.9% of Talk page activity, includes references to official Wikipedia guidelines, as guidance for article editing. This pattern generally followed "serious disagreements or flame wars" in response to high levels of conflict [26:8].

JUDGING THE CREDIBILITY OF INFORMATION ONLINE

A number of studies have examined factors impacting reliability perceptions of online sources, including Wikipedia. For example, Fogg et al. [4] studied over 2600 participants evaluating 100 real-world websites in 10 categories, identifying 18 top features that people consider when evaluating Web site credibility. Wikipedia fixes some of these to be the same across all articles, such as Design Look, Structure, Advertising, and Site Functionality. Other of Fogg's factors vary between articles, such as perceived information accuracy and bias, tone of writing, author motive, and readability; the discussion behind an article sometimes reveals signals about these factors. We hold most of these factors constant to focus specifically on how the presence and type of discussion influences perceptions.

Lucassen and Schraagen [16] build on this work with an experiment to discover "which elements of Wikipedia articles university students use to assess their trustworthiness." They found the major elements to be textual features such as comprehensiveness, correctness, length, and pictures. We hold these factors constant and extend the investigation to another factor, examining how the discussion behind content impacts trustworthiness evaluations when it is used.

Chesney [3] further suggests that people will rate content they are more familiar with as more credible than content on a random topic, which we observe but control for by random assignment.

Stvilia et al. studies a more objective information quality measure, noting early on that "The same information can be judged as being of different quality depending on the context of a particular use" [23:983]. This work describes a correlation between "actual" quality as measured by Featured Article status and discussion pages that are large, readable, and well-organized, but that translation of interest to this quality measure depends on the content of the discussion and whether or not a consensus has been reached [23:992]. Differences in collaboration patterns are known to lead to differences in actual quality of an article as measured by "Featured" status [9,15]. We would like to know if changing what somebody sees about the discussion correspondingly changes their perception of the article quality. Perceived quality is important for establishing legitimacy and building a more active community.

Stvilia et al. [23] call for empirical studies of information quality, suggesting the English Wikipedia as a particularly interesting case for study. We answer that call and provide empirical evidence for some of the paper's observations, noted above. That paper partially defines information quality as "noncontroversial," assuming greater capacity to objectively evaluate articles. Our extension to controversial topics serves as a building block for online deliberation systems that may eventually help people solve complex problems where objective evaluation may be impossible or less important than perceived quality.

The way that discussion around information is presented to users is an important design choice for an online community, with important consequences for how accurately the community is able to judge the credibility of information shared there. It may be, for example, that hoaxes are less likely to be discovered when discussion is separated from content, as in Wikipedia, compared to sites where discussion is front and center, such as Reddit, and that hoaxes are more likely to propagate in systems where the discussion is decentralized (as in Facebook as compared to Wikipedia or Reddit) [1,5]. It is important to understand how the visibility of discussion may change perceptions of quality in the system, and how the effects of salient discussion may depend on the content of that discussion.

A number of experiments in adding visible discussion capabilities to sites that previously did not support this kind of interaction around specific content have been tried and shown to add value. For example, Kriplean et al. [13,14] focus on grounded discussion and active listening, discussing the content even on sites that are already forums, as a means of increasing empathy and positive participation while efficiently summarizing and clarifying content.

Kittur, Suh, and Chi [10] cite significant distrust in user-generated content in current large collaborative systems, and hypothesize that readers lack sufficient information to evaluate the trustworthiness of content. They present a "Wikipedia dashboard" with visualizations of churn, reversion, and editor registration, showing who edited how much when, and found that the information raised or lowered trust ratings, depending on whether the record implied an irresponsible (e.g., many anonymous edits) or responsible process. Shneiderman [20] recommends full disclosure of sources' past performance patterns "in comprehensible and compact terms" as a means of building trust. Pirolli, Wollny, and Suh [19] extended that work with a more comprehensive view of edit histories, displaying only real data for each article in their lab experiment. They also demonstrated significant increases in article credibility ratings through exposure of editor identity and more detailed histories.

We build on this line of work by explicitly examining how readers' opinions of quality and trustworthiness may differ based only on what is exposed about the discussion that led to the content, termed content transparency [22]. In particular, we addressed the following research questions:

RQ1: What is the effect of exposing discussions about article content on perceived article quality?

Since discussion about controversial topics can reveal disagreements, inaccuracies, and possible bias, or remind readers that fallible non-expert editors created the content, revealing discussions could significantly diminish perceptions of quality (the "sausage hypothesis"). However, seeing the discussion could also show that content is being discussed and vetted at some level, which might increase perceptions of quality.

RQ2: Do different kinds of conflict resolution have different effects on perceptions of content quality?

Some aspects of the discussion, such as inequality in editors' experience levels, are known to lead to differences in rated quality, while other aspects, such as inequality of contributions, do not [2]. We investigate how differences between discussion strategies also influences rated quality.

Resolving conflict by personal attacks, threats to leave the community, or ignoring complaints seem likely to have much more of a negative effect than resolution accomplished more rationally, for example by citing policies or sources. We expected that conflicts resolved through collaboration would be seen as more likely to increase the quality of the resulting article text than other approaches to conflict resolution.

Montoya-Weiss, Massey, and Song [18] provide some hypotheses for how conflict resolution strategies in internationally distributed groups communicating asynchronously through text affect actual performance on a marketing consulting project with some time pressure. They found that avoidance and compromise behavior (as experienced by team members) significantly hurts performance, accommodation has no effect (because the text channel may not have been expressive enough to make this strategy obvious for team members to experience), and competition & collaboration correlate significantly and positively with performance. In [18], the compromise approach may have hurt performance because of its manifestation, cutting and pasting possibly contradictory content from different team members into a final team document, without integration effort.

RQ3: What do participants believe about how viewing the discussion may have changed their perceptions?

While RQ1 and RQ2 compare quality ratings from large groups of people who see different types of discussion or no discussion at all, RQ3 examines how individuals believe they are affected. In addition to their ratings of this particular article, we would like to see how the presence of discussion impacts participants' perceptions of Wikipedia overall.

We seek to discover whether the effects of different types of discussions are brought about by a deliberative process (what Kahneman calls System 2) or by a more automatic associative process (Kahneman's System 1) [7], and how these two mental systems interact to evaluate perceived quality. People are generally able to report with reasonable accuracy on System 2 activities, while associative activities generally escape awareness. A common experimental approach used to investigate especially System 1 psychology is to perform a between-subjects controlled experiment which demonstrates reliable differences caused by the manipulated variable, but where the cause of the effect never enters conscious thought (System 2), and participants may believe they were not affected or affected in the opposite direction. If we observe non-alignment between the results of RQ3 and RQ1 or RQ2, we know that at least part of the explanation must lie in what System 1 processes without the involvement of System 2.

It is possible to change the way people cognitively evaluate information, and Kahneman presents several factors known to more readily engage System 2. Simply presenting arguments has been shown to increase participants' cognitive evaluation of case descriptions in experiments evaluating the perceived legitimacy of Supreme Court decisions [17]. We extend that work by examining a source with a very different base level of credibility, examining how presentation of arguments (of different styles) may change cognitive evaluation and perceptions of quality.

EXPERIMENTS

Overview

We showed participants a segment of a collaboratively produced article and then showed them either no discussion or one of ten different discussions about the article segment. We then asked comprehension questions about the article and discussion to promote deep reading. We then asked participants to rate the article and the discussion on a number of scales, including ratings of article quality and the perceived level of conflict in the discussion. We collected data about demographics prior to the task, and collected self-report data about how the participants thought they may have been affected by seeing the discussion at the end.

Data Source

We chose Wikipedia as a ready source of data that has both a wide range of content along with publicly available discussion and resolution of disputes about the content. We selected topics that were both controversial and had high-quality articles. We wanted controversial topics with many disputes so we could identify a number of instances of different kinds of resolution. We also wanted articles that were high quality, because low quality articles are more likely to have features visible in the writing — poor style, lack of clarity, etc. — which could dominate judgments about quality, giving our manipulations less of a chance to have observable effects.

To identify controversial topics, we looked at community-curated lists of controversial issues and pages which were explicitly tagged as controversial, displaying and linking to an appropriate notice. We only considered articles in the main project namespace. This excludes, for example, Wikipedia policies and coordination pages, and uploaded files. This process identified 3403 unique controversial articles.

We then looked for examples of articles where the collectively generated content had reached a high level of quality. As with "controversial," we looked to the Wikipedia community's identification of the highest quality articles. The community uses an assessment scale to measure article quality, on an ongoing basis through a manual process. We selected Featured Articles, which are "the best articles Wikipedia has to offer, as determined by Wikipedia's editors." [32]. This measure correlates significantly with quality assessed by external raters [9]. At the time when we copied the list, there were 3189 featured articles (<0.1%). 50 articles were listed as both "controversial" and "featured," and we selected those for further investigation.

We examined this group of 50 articles and found common topical categories: Health, Science, Religion, Politics & History, Pop Culture, and Places. We chose articles across these categories, to ensure diversity in the pool of topics. We did our best to choose topics that were not currently dominating discussions in the news, and would be unlikely to change much over the course of the study period, because news reporting during the experiment could have unpredictable effects on the results. Our selections were Autism, Pope Pius XII, Yasser Arafat, and the Cretaceous–Paleogene extinction event.

One author read through the Talk page archives of these articles to determine the specific controversies present in that article. Some issues came up many times in different Talk page discussions, and the discussions were manually open-coded for the primary topic of discussion. Then, an issue was chosen based on its presence as an important controversy in that article, but which would be unlikely to be affected by strongly charged and very diverse viewpoints of our participants. As an example, in the Yasser Arafat article, we chose the controversy about his place and date of birth (a common controversy for celebrities) rather than the ones around his sexual orientation or the use of the word "terrorist." We did this to reduce variance and maintain a relatively constant level of emotionality, since high emotionality in conflict has been shown to lead to lower quality in conflict resolution outcomes [6].

We then created ten brief vignettes illustrating Talk page discussions about the particular chosen controversy, based in part on the styles (and some original content) from the discussions we observed. We portrayed a conflict between editors and resolution through each of the five approaches described by Thomas [24]. Two pilots of this study included only conflict conditions, and found that revealing strong conflict among editors lowered perceptions of article quality. To investigate rating differences beyond those that may be caused by seeing conflict, we also composed four non-conflict discussion conditions, three of which were based on the three most common Talk page uses described by Viégas et al. [26]. The fourth shows one editor reporting on changes s/he made to the article, with no apparent controversy; a second editor leaves one word of thanks. Also based on Viégas et al. [26], we had a sixth conflict condition depicting a single editor changing an article after removing a source that was inaccurately cited, implicitly addressing Wikipedia's policies around reliable sources; the editor is a frequent Wikipedian who also uses an abbreviation "POV" derived from the abbreviation of the "Neutral Point of View" core content policy [28]. We classified this as a conflict condition based on Viégas's description that this strategy is used as a response to conflict (p. 8). A manipulation check later verified that this condition "behaved" like a conflict condition with respect to the measure of perceived conflict.

These ten vignettes were originally created for one article and then adapted for each of the other topics by substituting in the topics, sides, and sources used in discussion, aiming to maintain similarity of structure and conversational style.

Some of the discussions contained excerpts from actual interactions, though all the discussions presented were created by one of the authors to represent a certain discussion type. These discussion vignettes were written with a number of principles in mind. All were written to be short to minimize the time and attention required of participants. The vignettes were written to clearly demonstrate each of the discussion types, and were similar across article topics (within the same discussion type). Because quotes were not direct, editors' names were replaced by two-letter initials, and each topic showed discussions among nominally different editors. Links and other aspects of Wikipedia formatting were retained in the presentation of both the article and discussion segments. Links to references were retained, but the actual references were not shown, to focus participants' evaluations on the text itself. No pictures were included in the short segments of our experiment. The actual discussions used are available in the attached appendix A.

Experimental Design: Details

As suggested by Kittur et al. [8], we posted a request on Amazon's Mechanical Turk requesting participation in our study and describing the task. Turkers who accepted our task were randomly assigned to an article topic and discussion type, neither of which they had seen before. We restricted eligibility to Turkers who had at least 95% of their prior micro-tasks approved, and who were in the United States (according to their Turk profile).

In all conditions, participants were first asked some preliminary questions about their background and use of Wikipedia. They were then shown a brief segment of an article, all participants assigned to a given topic viewing the same article text. Then they were shown a discussion about the article, and the discussion type varied as an experimental manipulation. We had ten discussion types, six of which displayed conflict and four of which did not, and an 11th (control) condition where no discussion was shown. We asked comprehension questions to provide greater motivation for participants to read the passages for understanding. The passages were presented as screenshots to exactly maintain Wikipedia formatting and presentation, and prevent participants from using the browser's Find or Copy-Paste features to answer the comprehension questions.

Two pilot versions of this study (with about 500 participants each) included an extra no-discussion control condition with extra article text to equalize the amount of time a participant spends on task in the control and the discussion conditions. We found no significant differences between the controls, so we dropped the extended text condition.

Among other questions, we then asked participants to evaluate the article using seven-point Likert items, each with {{Strongly, Moderately, Slightly} {Agree, Disagree}}, Neutral, and "I don't know" options. They evaluated the article with four questions:

"Based on the excerpt shown above, I believe the article as a whole is likely to be...

… well-written."

… an accurate and trustworthy source of information."

… biased." [Scale reversed for analysis.]

"Based on the excerpt shown above, I believe this article should be included in a collection of high quality articles."

The first three questions used for assessment of perceived article quality are also used in the Wikipedia Article Feedback Tool [27] version 4, which asked readers at the end of each article to "rate this page" on a scale of one to five stars on whether the article was "Trustworthy" (tooltip interpretations focused on the page having more or less reputable sources), "Objective" (tooltip interpretations ranged from "heavily biased" to "completely unbiased"), "Complete," and "Well-written." We adopt three of these same measures, omitting "complete" because we are showing participants only a small segment of the article. The wording about inclusion in a collection of quality articles is inspired by the similar dependent measure in [10].

Wikipedia's assessment tool also includes a checkbox allowing users to identify themselves as "highly knowledgeable" about an article's topic; we include this as a pre-task seven-point Likert item.

Turkers were given the option of participating again so long as we could guarantee that they would not see a discussion type or article topic more than once; in this experiment they were limited by the number of unique topics (four).

RESULTS

Participants

A total of 1348 surveys were completed by 566 unique Turk workers over the course of 26 hours. We paid $0.60 for each survey and a bonus of $0.15 to the 196 participants who participated the personal maximum of four times.

37 the 1348 surveys, completed by 15 participants, did not meet our *a priori* inclusion criteria that the worker must be in the United States (as discovered by a GeoIP lookup resolving outside the US) and those data were discarded.

Our participants were reasonably well-educated, as self-reported on a seven point scale, with 90% reporting at least some college experience, nearly a third reporting a bachelor's degree, and 16% reporting a post-graduate degree. They are also regular Wikipedia users, with over 80% reporting that they read Wikipedia "a few times per week" or more often, and over 96% reading "a few times per month" or more often. Over 85% agreed with the statement "Wikipedia articles are an accurate and trustworthy source of information." All of these factors were randomly distributed across the different experimental conditions.

A free-text box at the end, the only item explicitly labeled "optional," was filled out on 42.7% of surveys, showing that the Turkers were engaged enough to do extra work for us though it was not required for payment. 59.6% of the comments were substantive (beyond e.g. "no comments"). From this feedback, we learned in participants' words that the task was enjoyable, that it taught some of them to pay attention to their information sources, and that the pay was fair. These survey subsets were not significantly different on rated quality or on how participants thought they had been affected by reading the discussion.

To check for responses that may have been completed too hastily, we manually reviewed all responses that were completed in under three minutes, with that cutoff being slightly above one standard deviation below the mean. The comprehension question responses and overall response patterns were very similar to the remaining data, so we found no reason to exclude any other participants.

Measures

We examined the four measures of article quality (well-written, unbiased, accurate and trustworthy, and should be included in a collection of high quality articles) to see whether they did indeed all measure a similar construct (overall "quality") or whether they were independent dimensions. We found that they were significantly correlated (average $\rho = .570$; $p < .0005$), and loaded onto a single factor in a Principal Components Analysis, consistent with pilot results. Confident that there was only one underlying dimension, we then considered only the final (overall) quality question, as an interpretable measure of overall perceived quality for our primary dependent variable.

As a manipulation check, we asked participants whether a conflict was present in the discussion. Those in a conflict condition agreed fairly strongly (mean 1.78 Likert points, 95% CI [1.67, 1.88]) and those who weren't disagreed (mean −.65, 95% CI [−.83, −.46]); the difference was highly significant ($p<.0005$). The Avoidance and Competition resolutions were seen as having more conflict than the other conflict types, and referring to policies in removing a source had less, but there was a definite gap between the 95% confidence intervals for the mean measure of conflict in conflict and non-conflict conditions. These data provide evidence that the manipulation was strongly successful.

Of 1200 responses, most agreed with the pre-test statement "Anybody can edit Wikipedia," with the modal response (39%) at "strongly agree;" only 10.3% disagreed with that statement and an additional 4.5% were neutral. However, participants generally did not use Talk pages. When asked how often they read Talk pages, 23.6% said "I don't know what these are" and an additional 44.6% of the 1303 responses said "never." 20.4% read Talk pages "a few times per year" and only 11.4% read them more often.

Analysis

Where means are reported in Likert points, they are on a seven point scale coded for analysis in the range [−3,+3], and otherwise unaltered from the participants' responses. To compare subsets of the data and see if viewers rated the same article text as being of different quality based on the style of discussion they saw, we report the results of a Kruskal-Wallis one-way analysis of variance, which is a nonparametric test that permits ordinal Likert item data. Mean numeric values are reported to allow the reader to estimate effect size.

In order to ensure that our results were valid under strict statistical assumptions regarding fully independent measurements and rule out order effects, we repeated our analysis on just the first survey that each participant loaded, and identically significant effects were observed. We also replicated the analysis using ANOVA instead of Kruskal-Wallis, and again observed the same effects. Treating the data either as ordinal (Kruskal-Wallis) or interval (ANOVA), the results do not vary.

By randomly assigning participants to experimental conditions, we expected prior knowledge and perceptions about Wikipedia, the topic, and our task to be randomly distributed across conditions. We have verified that prior knowledge and trust of Wikipedia are not significantly different between the groups that we are comparing below.

RQ1: How does exposing discussions about article content affect perceived article quality?

Our first planned contrast examined whether or not exposure to the discussion (of any type) led readers to assess the article at a different quality level than readers who did not see the discussion behind it. We find that *people who saw any Talk page discussion rated quality significantly lower than those who did not see any discussion* (0.21 vs. 1.08 Likert points, p < .0005).

We hypothesize that if people are exposed to conflict about an issue, they would be less likely to perceive a discussion outcome as high quality. We find that *people who saw conflict conditions rated quality lower than those who saw non-conflict conditions* (−0.04 vs. 0.61 Likert points, p<.0005). We also compared quality ratings of those who saw non-conflict discussions with those in the control condition, and found that *even those who saw non-conflict conditions rated quality lower than those who did not see any discussion* (0.61 vs. 1.08 Likert points, p=.011).

RQ2: Do different kinds of conflict resolution have different effects on perceptions of content quality?

After finding the main differences between no discussion, non-conflict discussion, and conflict discussion conditions, we looked within these groups to see if particular strategies for conflict resolution led to higher or lower quality ratings than others.

We found no significant differences between quality ratings among non-conflict discussions; none were expected.

Based on the prior literature (e.g. [18]) and pilot data, we hypothesized that the “ignored complainer” in the “avoidance” strategy would lead to significantly lower quality ratings than the other discussion types, because it shows one editor complaining about low quality and nobody responding to those criticisms. Our data support this hypothesis, whether comparing to just the conflict conditions (−1.13 vs. 0.16 Likert points, p<.0005) or to all discussion conditions (−1.13 vs. 0.36 Likert points, p<.0005).

Based on prior literature that describes collaboration as the conflict resolution strategy generally leading to the best outcomes, we believed the collaboration conflict resolution strategy would be perceived as a “good” conflict resolution strategy and enhance output quality ratings, at least in comparison to other ways of resolving conflicts. We hypothesized that the collaboration condition would lead to significantly higher perceived quality than the other conflict types. Our data support this: People who saw the collaboration condition rated quality higher than those who saw other conflict conditions (0.61 vs. −0.18 Likert points, p<.0005). People who saw the compromise condition also rated quality higher than those who saw other conflict conditions (0.47 vs. −0.13 Likert points, p=.004). The compromise and collaboration conditions were not distinct from each other; they were also illustrated similarly. In fact, these two were not significantly different from the non-conflict discussion conditions in terms of quality rating, even though they were significantly different from the non-conflict discussion conditions in terms of perception of conflict (1.70 vs. −0.65 Likert points, p<.0005). These two resolution strategies were able to overcome the negative effects of conflict on quality ratings.

RQ3: What do participants believe about how viewing the discussion may have changed their perceptions?

Participants who saw discussion conditions were asked directly if reading the discussion affected their perceptions, and if so in what direction. The question, options, and data for each discussion condition are shown in Figure 2. Participants believed that seeing the discussion *raised* their perception of the article’s quality and of Wikipedia in general, overall and for many single-discussion-condition subsets, with significant effects indicated by asterisks in the figure. The ignored complainer in the “avoidance” condition is the only one where participants were aware of the fact that reading the discussion lowered their perception of that article’s quality.

Our exact significance test in these analyses examined the balance between the “raised” response options as one set and the “lowered” response options as another set, calculating how likely the actual balance between these would be if both were equally likely, akin to calculating the probability that at least a given percentage of “heads” would appear on the appropriate number of flips of a fair coin. Unless otherwise indicated, significant effects have p<.0005 and non-significant effects have p>.05.

CONCLUSIONS AND DISCUSSION

Large-scale distributed systems for collaborative content creation must have a good way of dealing with controversial topics. We have shown that the way these conflicts are dealt with, and whether that is exposed to consumers of the content, impacts how viewers perceive the quality of the content, even when its actual content is held constant.

We find evidence that surfacing discussions about content generally lowers the perceived quality of the content, and that this effect differs significantly depending on how the discussion is conducted and whether or not a conflict is revealed. However, the effect runs counter to participants’

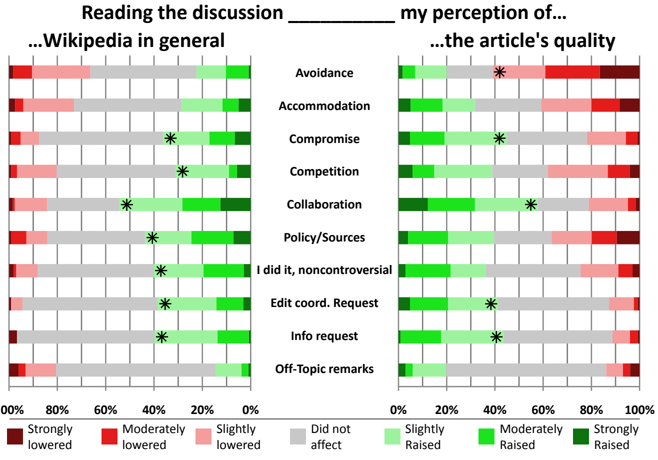

Figure 2: How participants think viewing the discussion affected their perceptions

self-reported perceptions, as participants tended to report that reading the discussion increased their perception of the article's quality and of the overall platform (here, Wikipedia) in general. The discussion which follows explores possible mechanisms that may help explain the effect.

We suspected that the quality ratings of those already familiar with Talk pages would not be affected by seeing discussion (especially non-conflict discussion) in the same way as somebody who had never seen Talk pages, because the discussion may have made the less Wikipedia-savvy people more aware of the fallibility of Wikipedia's editing process. If the presence of discussion changes how people engage with the article in a way that reduces perceived quality even among individuals already familiar with discussion pages, and the same results are observed in that subset, we would reject that explanation. To test for this, we asked at the beginning of the exercise how often participants read Wikipedia Talk pages, with results summarized above, and filtered to use only responses where people reported familiarity with Talk pages. While some statistical tests suggested the same primary results, when using non-parametric ordinal tests and strictly maintaining the limit of one observation per Turker, we did not have sufficient power to observe significant effects that would lead us to conclusively eliminate this possibility.

Our experimental design, across all conditions, aimed to engage participants in reasonably deep thinking. Our comprehension questions required all participants to read in the material in a reasonable degree of detail; our presentation of those passages as images helped force more detailed reading and engagement.

However, seeing the discussion may have engaged participants' critical thinking skills (Kahneman's System 2) more deeply than the non-discussion condition, causing them to look at the original material more skeptically, leading participants to be more critical of the article in their ratings. Interestingly, they nevertheless believed the discussion caused them to perceive the article as higher quality, perhaps because they believe they were able to make a more accurate and informed evaluation, or because System 2 included a more thoughtful understanding of "quality" than the quick intuitive judgment of System 1, which often substitutes easier questions that may have different baseline answers [7]. If this is the explanation for our results, it would suggest that designers of online collaborative content creation systems need to pay close attention to factors that influence System 2 engagement (summarized in [7]) when considering how people will perceive content quality.

As another possible explanation, we see from the design literature that people are less willing to be critical of work that they perceive to be "finished" or "complete," and more willing to offer criticism of works in progress. People are more willing to give higher-level constructive criticism about something that seems to be more "sketchy" with lower level details not yet fixed, as compared to a polished product [31]. This may be a mechanism of System 1.

All our article segments were taken from works currently listed as Featured Articles, meaning they have already reached the highest quality standard on Wikipedia, but all Wikipedia articles are in some sense incomplete [29]. Revealing the discussion might frame the article more as a "work in progress" than a completed, polished piece. This more apparent state of incompleteness could invite more criticism, as reflected in lower quality ratings, even while participants feel that reading the discussion improved their perceptions (presumably compared to other work viewed as being in a similar state of completion). We plan to check for this in a future study that manipulates the perceived completeness of collaboratively produced content, regardless of the presence or absence of discussion. If this explains our experimental results, system designers would need to attend to stylistic details that make work seem more or less polished, depending on their goals for the system.

This paper describes an experiment and a robust set of results about *how* the presence of discussion causally affects perceived quality. The results prompt questions of exactly *why* these results are observed, which must be explored in future work, before specific prescriptive advice can be given to the designers of online community platforms. The causal understanding will also have implications for designers of external tools that reveal information about the editing and production process, such as many research tools visualizing Wikipedia edit histories with the intention of impacting perceptions of trustworthiness, accuracy, quality, and community (e.g. [10, 19, 25]).

FUTURE WORK

In this work, we have found a "sausage" effect that revealing discussion generally lowers perceived article quality, with the strength of the effect depending on the presence of conflict and the way that any present conflict is resolved. This leaves open many questions about the underlying causal psychological mechanisms.

Since we studied only one context, it naturally leads one to wonder about the generality of this finding. One could readily use a similar experimental paradigm with other sources or in contexts with different signals about information quality, for example. The presence and type of discussion seems likely to interact with other factors impacting a source's credibility, such as assumed editor expertise (e.g. showing editorial discussions from a major national newspaper, or a federal agency's rule-setting work, instead of Wikipedia) or page presentation (e.g. does showing reasoned discussion behind content on a site that does not otherwise appear to be a credible source improve credibility?).

If conflict and conflict resolution strategies in discussions can be automatically detected (e.g. by a machine learning classifier), one could (a) replicate some of these findings on a larger scale, using ratings from the Wikipedia article assessment tool, and (b) proactively flag certain future discussions for moderator attention and possible intervention.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from NSF grants IIS-1111750, SMA-1064209, OCI-0943168, IIS-1149797, IIS-1217559, IIS-0968484, OCI-0943148, IIS-1111124, CNS-1040801, and from the Center for the Future of Work, Heinz College, Carnegie Mellon University. We also thank our participants and our reviewers for some very high-quality reviews.

REFERENCES

1. Y. Appelbaum. “How the Professor Who Fooled Wikipedia Got Caught by Reddit.” *The Atlantic*, 2012. http://www.theatlantic.com/technology/archive/2012/05/how-the-professor-who-fooled-wikipedia-got-caught-by-reddit/257134/.
2. O. Arazy and O. Nov. “Determinants of Wikipedia Quality: The Roles of Global and Local Contribution Inequality.” *Proc. CSCW*, ACM (2010), 233–236.
3. T. Chesney. “An Empirical Examination of Wikipedia’s Credibility.” *First Monday 11*, 11-6 (2006).
4. B.J. Fogg, C. Soohoo, D.R. Danielson, L. Marable, J. Stanford, and E.R. Tauber. “How Do Users Evaluate the Credibility of Web Sites?” *Proc. DUX*, (2003), 1.
5. M. Garber. “Abraham Lincoln Did Not Invent Facebook: How a Guy and His Blog Fooled the Whole Wide Internet.” *The Atlantic*, 2012. http://www.theatlantic.com/technology/archive/2012/05/abraham-lincoln-did-not-invent-facebook-how-a-guy-and-his-blog-fooled-the-whole-wide-internet/256945/.
6. K.A Jehn. “A Qualitative Analysis of Conflict Types and Dimensions in Organizational Groups.” *Administrative Science Quarterly 42*, 3 (1997), 530–557.
7. D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
8. A. Kittur, E.H. Chi, and B. Suh. “Crowdsourcing User Studies with Mechanical Turk.” *Proc. CHI*, ACM (2008), 453–456.
9. A. Kittur and R.E. Kraut. “Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination.” *Proc. CSCW*, ACM (2008), 37–46.
10. A. Kittur, B. Suh, and E.H. Chi. “Can You Ever Trust a Wiki?” *Proc. CSCW*, ACM (2008), 477.
11. A. Kittur, B. Suh, B.A. Pendleton, and E.H. Chi. “He Says, She Says: Conflict and Coordination in Wikipedia.” *Proc. CHI*, (2007), 453–462.
12. M. Klein and S.C.-Y Lu. “Conflict Resolution in Cooperative Design.” *Artificial Intelligence in Engineering 4*, 4 (1989), 168–180.
13. T. Kriplean, M. Toomim, J. Morgan, A. Borning, and A.J. Ko. “Is This What You Meant?: Promoting Listening on the Web with Reflect.” *Proc. CHI*, ACM (2012), 1559–1568.
14. T. Kriplean, M. Toomim, J.T. Morgan, A. Borning, and A.J. Ko. “REFLECT: Supporting Active Listening and Grounding on the Web through Restatement.” *Proc. CSCW*, ACM (2011).
15. J. Liu and S. Ram. “Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality.” *ACM Trans. Manage. Inf. Syst. 2*, 2 (2011), 11:1–11:23.
16. T. Lucassen and J.M. Schraagen. “Trust in Wikipedia: How Users Trust Information From an Unknown Source.” *Proc. 4th Workshop on Information Credibility*, (2010), 19–26.
17. J.J. Mondak. “Perceived Legitimacy of Supreme Court Decisions: Three Functions of Source Credibility.” *Political Behavior 12*, 4 (1990), 363–384.
18. M.M. Montoya-Weiss, A.P. Massey, and M. Song. “Getting It Together: Temporal Coordination and Conflict Management in Global Virtual Teams.” *The Academy of Management Journal 44*, 6 (2001), 1251–1262.
19. P. Pirolli, E. Wollny, and B. Suh. “So You Know You’re Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility.” *Proc. CHI*, ACM (2009), 1505.
20. B. Shneiderman. “Designing Trust into Online Experiences.” *Communications of the ACM 43*, 12 (2000), 57–59.
21. H.A. Simon. *The Sciences of the Artificial*. MIT Press, 1996.
22. H.C. Stuart, L. Dabbish, S. Kiesler, P. Kinnaird, and R. Kang. “Social Transparency in Networked Information Exchange: A Theoretical Framework.” *Proc. CSCW*, ACM (2012), 451–460.
23. B. Stvilia, M.B. Twidale, L.C. Smith, and L. Gasser. “Information Quality Work Organization in Wikipedia.” *Journal of the American Society for Information Science and Technology 59*, 6 (2008), 983–1001.
24. Thomas, K.W. “Conflict and Conflict Management: Reflections and Update.” *Journal of Organizational Behavior 13*, 3 (1992), 265–274.
25. Viégas, F.B., Wattenberg, M., and Dave, K. “Studying Cooperation and Conflict Between Authors with History Flow Visualizations.” *Proc. CHI*, ACM (2004), 575–582.
26. Viégas, F.B., Wattenberg, M., Kriss, J., and van Ham, F. “Talk Before You Type: Coordination in Wikipedia.” *HICSS*, (2007), 78.

27. Wikimedia Foundation. "MediaWiki Extension:ArticleFeedback." https://www.mediawiki.org/wiki/Extension:ArticleFeedback.

28. Wikipedia community consensus. "Wikipedia: Neutral Point Of View." *Wikipedia*, 2012. https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=491462094.

29. Wikipedia community consensus. "Wikipedia: There Is No Deadline." *Wikipedia*, 2012. https://en.wikipedia.org/w/index.php?title=Wikipedia:There_is_no_deadline&oldid=487857034.

30. Wikipedia contributors. "Help: Using Talk Pages." *Wikipedia*, 2012. https://en.wikipedia.org/w/index.php?title=Help:Using_talk_pages&oldid=487941001.

31. Wong, Y.Y. "Rough and Ready Prototypes: Lessons from Graphic Design." *Posters and short talks of CHI*, ACM (1992), 83–84.

32. *Daily Cleveland Herald*, 1869.