Conflict in Comments: Learning but Lowering Perceptions, With Limits

W. Ben Towne, Carolyn P. Rosé, & James D. Herbsleb
Carnegie Mellon University
Pittsburgh, Pennsylvania
{wbt, cprose, jdh}@cs.cmu.edu

ABSTRACT
Prior work and perception theory suggests that when exposed to discussion related to a particular piece of crowdsourced text content, readers generally perceive that content to be of lower quality than readers who do not see those comments, and that the effect is stronger if the comments display conflict. This paper presents a controlled experiment with over 1000 participants testing to see if this effect carries over to other documents from the same platform, including those with similar content or by the same author. Although we do generally find that perceived quality of the commented-on document is affected, effects do not carry over to the second item and readers are able to judge the second in isolation from the comment on the first. We confirm a prior finding about the negative effects conflict can have on perceived quality but note that readers report learning more from constructive conflict comments.

Author Keywords
Experiment; validation; social influence; comments; crowdsourcing; distributed evaluation; creative work.

ACM Classification Keywords
H.1.2. User/Machine Systems: Human factors, Human information processing; H.5.2 Evaluation/methodology, Natural language, Interaction styles; H.5.3 Group and Organization Interfaces: Computer-supported cooperative work, Web-based interaction, Evaluation/methodology.

INTRODUCTION

"Wicked problems" such as climate change may be more complex than any specific individual or committee is capable of completely understanding by themselves. However, evidence indicates that a large and distributed set of cognitive resources may be available to help solve such problems, if we can figure out how to use them well [22]. Addressing one set of approaches to large-scale distributed problem solving, a 2016 *Science* article by Michelucci & Dickinson asserts that human computation has "huge potential" to address wicked problems, going beyond microtask crowdsourcing to create shared online workspaces where participants "contribute, combine, revise, connect, evaluate, and integrate data and concepts" [18]. In many platforms designed to support large-scale distributed problem solving, basic ideas are crowdsourced and posted publicly, where others can evaluate the contribution and/or add comments and/or read the comments others have written. However, there are questions being raised in at least the public dialogue about the potential of spill-over effects from comments and interactions posted online to undermine the significant positive potential of Internet-based collaboration [24].

Where others' comments on or ratings of content are also visible, as is often the case, they can have an impact on the level of quality perceived by subsequent readers. Social influence on perceptions of quality could create a feedback loop that leads groups of people to irrationally herd toward large group evaluations that are path-dependent and not necessarily connected to the quality of what is being perceived [1,21]. Causes and consequences of market-based valuation bubbles and herding in asset pricing have been studied (e.g. [11]), but little is known about how much of that knowledge applies to distortions in perceptions of value when the signals and consequences are more purely social, in the absence of pricing or direct economic incentives.

Examining social influence effects in large online community platforms, Muchnik, Aral, & Taylor conducted an experiment in which fresh and unrated content, which users had posted to a social news aggregation Web site, was *randomly assigned* to an initial upvote or downvote (or control condition with neither). The random upvote increased the probability of up-voting by the first viewer by 32% without a corrective increase in down-voting; the effects of that initial manipulation persisted and increased mean ratings by 25% five months later. A random initial downvote doubled the probability of subsequent downvotes, but also significantly increased the probability of a subsequent corrective upvote. "Friends" of the commenter were more likely to upvote in response to a randomly assigned initial vote (up or down) than to a post with no random initial vote, but even when differences in the probability of voting were considered, there remained differences attributed to statistically significant opinion

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025902

change resulting from the random initial vote. Their results "suggest that social influence substantially biases rating dynamics in systems designed to harness collective intelligence." That paper concludes by calling for more research exploring mechanisms driving individual and aggregate ratings, especially in real social environments, as "essential to our ability to interpret collective judgment accurately and to avoid social influence bias in collective intelligence" [19].

In the setting of Wikipedia, Towne, Kittur, Kinnaird, & Herbsleb [26] showed that when readers are exposed to discussion behind collaboratively edited content, their assessment of the quality of the content is lower than readers who do not see the discussion. The effect is stronger if the discussion contains conflict, but this strengthening can be erased if the discussion shows a Compromise or Collaboration resolution strategy from the editors involved. The effect, observed in a between-subjects experiment, was in that study counterintuitively accompanied by participants' self-reports that reading the discussion *increased* their perceptions of article quality.

The present study seeks to test the extent of that prior work's finding, using a similar methodology in a different setting and measuring the extent to which this effect might extend beyond the article being discussed directly, to other articles that are similar along a couple selected dimensions.

METHOD OVERVIEW

In this study, we show participants a segment from each of two crowdsourced proposals entered into a platform designed to support large-scale collaboration around a complex issue, namely global climate change. Following the first proposal (only), most participants saw a comment associated with that first proposal. We experimentally varied the type of comment shown in a 3x2 full factorial design, plus a no-comment control, as described in "Experimental Conditions" below. We also independently varied the relationship between the first and second proposals (i.e. same author, similar topics, or neither) as described below. After reading each proposal (and if applicable, the comment immediately following the first proposal), participants were asked to evaluate the proposal quality with a multi-item scale. We look to see if those quality evaluations changed as a result of the experimental conditions to which participants were randomly assigned.

HYPOTHESES

Hypotheses in this work are presented here in descriptive form, summarized in Table 2. Based on work by Towne et al. [26], we hypothesize that (H1) the presence of a comment will negatively affect participants' ratings of the first proposal's quality, and (H2) the size of the effect will depend on the type of comment, such as whether it presents conflict. The size of the effect may also depend on how that conflict is resolved [26]. This study investigates further to determine if (H3) the effect extends to other proposals (i.e. the second one participants saw) which are similar or different in certain ways.

In investigating that possible spill-over effect, we experimentally manipulate whether comments are directed at the *content* or the *author.* We hypothesize that (H4) if the comment is attributed to something about the *content* (e.g. a claim that the main idea is significantly flawed), the effects of reading that comment might extend to other topically similar proposals regardless of who wrote them, and (H5) that if the comment is something attributable to the *author*, the effects might extend to other proposals by the same author regardless of proposal topic.

Whether or not H1 and H2 were supported, we wanted to be able to better understand the processes underlying any differences in ratings between experimental conditions. If the same set of processes is at work here as observed previously [26], we hypothesize (H6) that participants would report *beliefs* that their perceptions of quality had been affected in the *opposite* direction that the between-subjects analysis showed that it had been. To explore this hypothesis, we asked participants who saw comments about the degree to which they thought reading the comments raised/lowered their quality perceptions, as two questions on a page immediately following participants' evaluations.

We also asked participants to tell us whether or not they learned anything from the comments, to investigate a potential benefit they might provide and put H2 in context, as comments offering a different perspective may provide the reader with more novel and/or useful information than comments which do not express both a conflict and resolution. This allowed us to explore the hypothesis (H7) that participants would learn most from comments containing constructive conflict.

In all conditions, content was presented in the same order. We recognize that reading the first proposal could have an order-specific effect on the rating of the second, separate from underlying differences between the proposals that might lead to rating differences between the first and second proposals. This experiment specifically investigates one hypothesized order effect, namely how reading a comment and then providing an evaluation which may be influenced by it may also influence perceptions of material read immediately after it, on the same page. Having the rating materials on the same page increases the probability of observing (and being better able to characterize) a carry-over effect if one exists. Participants are randomly assigned to a comment condition (or control condition absent any comments), and the primary analysis compares between those conditions.

MATERIALS

Corpus

Inspired by examples of successful large-scale collaboration elsewhere, such as Wikipedia or FoldIt, Malone et al. created the MIT Climate CoLab to be a global platform for

collaboratively developing and evaluating proposals for what to do about global climate change [16]. In an annual series of contests, its members have collectively produced, commented on, and voted on over 1500 proposals, typically 1500-3000 words in length [4], on a wide range of climate-change-related topics. At the start of the main 2016 contests, the site had over 50,000 registered user accounts. We used the 1501 proposals in contests that were then completed as the documents in our model and experiment.

In recent years, several contests have been run in parallel, each addressing different subtopics (such as Transportation, Waste Management, Buildings, and Energy Supply) or calling for plans that integrate a number of other proposals [17]. Participants and problems discussed come from many countries, though most participants come from North America [6]. The interface, proposals meeting the language requirements [2], and all materials used in this experiment are in English.

We further restricted our materials to the 350 Climate CoLab proposals submitted in contests between 2011 and 2015 inclusive that made the semifinalist round as selected by expert judges. We chose semifinalist proposals (or higher) to establish a minimum degree of quality, similar to Towne et al.'s procedure for choosing high-quality articles in [26]. We set the threshold at semifinalist proposals as opposed to finalists or winners to maintain some quality while also ensuring that the pool of proposals to choose from would be sufficiently large (as in [20]).

We further eliminated proposals that could not be basically understood by the summary text, e.g. because the proposal relied heavily on reference to other proposals or documents. We removed two proposals whose content had been submitted in the form of images, and another whose latest version simply read "The new version will be posted shortly" because the text of these had not been properly captured for topic modeling.

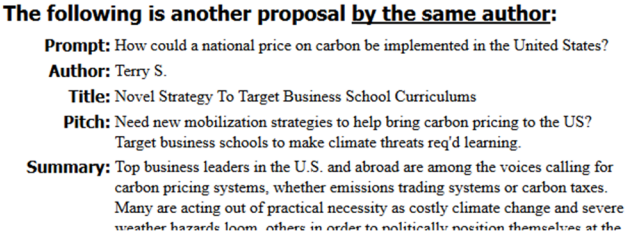
Experimental Conditions

Comments shown on the first proposal were randomly assigned from a 2x3 factorial design, plus a "no comment" control which helps us test H1. The three levels of the "conflict" factor (Conflict without resolution, Conflict with resolution, and Non-conflict) were chosen based on the categories distinguished by previous work [26], and help test H2. The two levels of the "direction" factor (content-directed vs. author-directed) help test H4 and H5.

The second proposal was chosen randomly from the cells in Table 1. We excluded the "Same author, similar content" cell because when observed, the proposal pairs in this cell are largely copies of each other, sometimes with little or no modification, and in allocating our budget for experimental conditions, it was least interesting to see if comments on one proposal would affect perceptions of another copy of that proposal. The relationship between proposals was called out in bold underlined header text just before the presentation of the second proposal, as shown in Figure 1.

4 similarity types (2x2):	*Same **A**uthor*	*__D__ifferent Author*
*Similar **C**ontent*	≈ Copies (not tested)	Relationship C
__D__issimilar Content	Relationship A	Relationship D

Table 1: Relationship second proposal has to first proposal, randomly assigned. This factor crosses author and content similarity factors, a factorial design excluding one cell.


The following is another proposal by the same author:

Prompt: How could a national price on carbon be implemented in the United States?

Author: Terry S.

Title: Novel Strategy To Target Business School Curriculums

Pitch: Need new mobilization strategies to help bring carbon pricing to the US? Target business schools to make climate threats req'd learning.

Summary: Top business leaders in the U.S. and abroad are among the voices calling for carbon pricing systems, whether emissions trading systems or carbon taxes. Many are acting out of practical necessity as costly climate change and severe

Figure 1: The relationship between proposals was explicitly called out. This screenshot shows Relationship A from Table 1. For Relationship C, the underlined text read "about a similar topic" and for Relationship D, it read "that was submitted" without underline.

Selecting focal proposals

As potential focal proposals, we considered those where other proposals existed that fit the "same author, unrelated content" criteria described below. We selected multiple focal proposals so that our experimental results would not be too sensitive to specific details of any particular proposal, but kept the number small so that we could collect relatively robust sets of quality ratings on each and analyze aggregated results while still controlling for any real quality differences that may exist between selected proposals.

In order to better cover the space of available proposals, we used a constraint satisfaction solver to maximize diversity by selecting the set of four focal proposals by four different authors that were on average maximally different from each other according to a previously studied LDA/cosine similarity measure [27] (the same as used in "Selecting topically related proposals" below). The solver we used was Excel 2013's "Evolutionary" solver, which produced better results than its "GRG nonlinear" solver.[1]

Selecting topically related proposals

For each focal proposal, we selected the proposal that had no overlap in author team and the greatest topical similarity to the focal proposal, operationalizing relationship "C" from Table 1. Topical similarity was measured by the cosine similarity of proposals' topic vectors according to a

[1] For more detailed citations, see http://www.solver.com/excel-solver-algorithms-and-methods-used.

Latent Dirichlet Allocation (LDA) model run over the entire corpus with program default hyperparameters.

Consistent with [27] and with Xu and Ma's goal of maximizing dissimilarity between clusters [31:303], we selected the number of topics by choosing the model with the lowest average percentage of proposals that has neither or both topics in each possible topic pair. The model was run to 1000 iterations for coarse tuning between 5 and 300 topics, and 2000 iterations for fine tuning between 50 and 60 topics, concluding with a 57-topic model, as the one which maximally separated proposals into different topics.

In a prior experiment, Towne et al. [27] found this measure to match human perceptions of which pair among three documents were most similar two-thirds to three-quarters of the time. Using this measure of topical similarity instead of the CoLab contest categories helps us generalize results beyond the Climate CoLab structure which requires manual creation of a topic hierarchy and manual assignment of proposals to those categories (here, by proposal authors who are not experts about the categorization scheme).

Selecting topically unrelated proposals

For each focal proposal, we sorted other proposals according to the same LDA/cosine similarity scale and randomly selected from the bottom half of that distribution, ensuring that it was by a different author, operationalizing relationship "D" referenced in Table 1.

We drew proposals operationalizing relationship "A" in Table 1 from this same bottom half of the distribution. If the author had multiple proposals there, we chose the one more different from the focal proposal (this happened in only one of the four sets, and the two options were adjacent to each other in the list). In all eight cases, these proposals were from contests different than the focal proposal.

Proposal Tweaks

Proposal summaries were modified from the originals for greater suitability for use in the experimental setting and consistency with each other in the following ways. Hyperlinks and text references to other proposals were removed, to make the summary more self-contained. Special formatting was removed. Acronyms that were not clearly introduced but discoverable in the original context (e.g. GHG) were spelled out. Spelling and grammar were cleaned up so that these attributes of writing would not dominate or interfere with perceptions of quality based on other factors, removing this potential source of variance for a cleaner experiment. One proposal summary was shortened (largely by removing redundancy) to bring its length in line with the others, appropriate for a Mechanical Turk task. After these modifications, the proposal summary lengths ranged from 168 to 350 words, with an average of 290 and standard deviation of 48. Material length was comparable with a previous study based on news articles [25].

Selecting names

Because author attribution is a factor being explored in this experiment, we attach author names to proposals to help underscore the "same author" relationship. However, researchers presenting experimental materials with names attached need to be aware of the impact of those names, because aspects of the name (such as perceived gender) can change the way participants perceive the materials attributed to them [8].

Through a set of four studies, Fleet and Atwater [8] identify four most gender-neutral *first/given names* and of these four we chose the two most prevalent ("Expected total number alive today" male + female total) in the United States according to the Wolfram|Alpha Knowledgebase, 2016 (e.g. [29]), which were Terry and Lee respectively.

We used only a surname initial instead of a full surname according to the advice of Kasof [14:152]. The US Census publishes the frequency of *surnames* occurring >100 times, covering 90% of the population [30]; at the time of study the most recent available data is from the 2000 Census. (First name information is not available from the 2000 census.) Within this data, the most common first letter (surname initial) is M, followed closely by S.

We crossed these two for balance. In our experiment, the focal proposal was always credited to "Terry S." and the other proposal, if by a different author, was credited to "Lee M." An experiment by Howard & Kerin [12] found that *similarity* between a participant's name and the putative name of an author whose work the participant was evaluating (in the experiment, first name and last initial were shown to match or not match participants'), engages self-referencing and increases thoughtful examination. Here, we did not ask for participants' real names, but we don't expect any rare overlaps to lead to differences between the randomly assigned experimental groups.

Dependent Measures

As in work by Towne et al. [26], we consider quality assessment criteria used in the community where the data comes from. In the Climate CoLab, proposals are assessed by expert judges along four scales: Feasibility (in four specific aspects), Novelty, Impact, and Presentation [3]. These criteria and their descriptions have been consistent through the history of the platform. We requested each participant evaluate each proposal using the following seven-point Likert items, each with {{Strongly, Moderately, Slightly} {Agree, Disagree}}, Neutral, and "I don't know" options:

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

...be {technically, economically, socially, politically} feasible." (4 questions, answers averaged for feasibility)

...be novel, reflecting innovative thinking and originality."

...make an impact on the issue raised in the prompt, if it were implemented."

...be well-presented."

For an overall measure of quality, on the same scale, we also asked included an item (from [15] and [26]) *"Based on the excerpt shown above, I believe this proposal should be included in a collection of high quality proposals."* This is the fifth item in a five-item perceived quality scale that is our primary dependent variable. In relatively rare cases where a participant refused to answer any particular item, that item simply carries no weight in computation of the composite (mean) rating for that participant and scale.

Demographics & prior expertise

Before the task, participants answered questions about their educational background and level of interest and experience in the general domain area of the content material, including questions copied from prior surveys of the community where materials came from.

We recognize that prior knowledge about a topic has been identified as a potentially dominant factor in quality evaluations [25] in work that explicitly seeks follow-up studies illuminating the impact of issue familiarity on the effect of comments [25:71]. Also, people who have more prior knowledge or are more involved do more elaborate information processing, attend to more quality cues, especially intrinsic ones, and may be less extreme but possibly faster in their overall quality judgments [23:315]. Therefore, after each evaluation, we included an item "I have expertise on the topic(s) discussed in this proposal." As the final multiple-choice question in our study, we asked participants about their level of familiarity with the platform from which materials were drawn.

Participants & Filters

We recruited participants on Amazon's Mechanical Turk paid crowdsourcing platform, in part because the population for generalizability of results is intended to broadly include those likely to visit/use crowdsourcing platforms. (Our materials come from a different crowdsourcing platform, described above.) This choice of participant pool also has several other benefits relevant to experiments exploring the impact of social information on perceptions in *CHI* [13]. Our tasks were posted midweek, on days describes as slow (with respect to the volume of tasks being posted) on the Turker Nation forum. Tasks were posted through afternoon and evening times and continued at a slower pace into the following day.

We reduced potential evaluation variance by restricting participants to those who were in the United States according to their Turk profile, and prior to analysis we filtered out any data from participants who did not have a GeoIP lookup resolving to the United States. We also restricted participation to those who had at least 500 assignments approved by other requesters and a 95% overall approval rating. (This is the same as in [27]). In the main experiment, we also excluded from analysis two participants who wrote keyboard-mashing strings in unselected "other" boxes on demographic questions, and two participants who took steps to defeat the participation limits. We filtered out those who completed the main experiment in under 2.5 minutes because less than one minute per proposal plus 30 seconds for demographic questions indicates those participants are less likely to have fully read the materials or questions.

Both usage goals and time pressure also affect perceptions of quality [23:316], and we held these constant across experimental conditions. During the task, we specified usage goals (proposal evaluation) and did not add any time pressure beyond what is self-imposed by participants independently of the task (similar to e.g. [25]), permitting an hour for a task which typically took several minutes.

At the very end of our task, participants also had an open text box to optionally provide feedback about what they had just completed. We have found this to be a best practice that can facilitate detection of certain errors in experimental setup if they exist, provide qualitative feedback that can help inform future analyses and/or task designs, and increase participant satisfaction by allowing participants to express any remaining thoughts they wanted to express. We read all feedback submitted.

MANIPULATION CHECK ON COMMENTS

Manipulation Check on Comments: Design

The comments for each focal proposal and experimental type were created by authors of this paper based on a review of real comments left on proposals on the site, similar to the setup in [26].

Before proceeding with the experiment, we checked to see if the differences between experimental conditions were manipulating the intended constructs. To do this, each of the 24 comments was posted to Turk along with the following 8 statements assessed on the same 7-point agree/disagree scale described above, followed by a free-text feedback field:

1. This comment is civil.

2. This comment is directed at the *author* of the proposal.

3. This comment is directed at the *content* of the proposal.

4. The author of the comment and the author of the proposal likely have some conflicting views, at least regarding what this comment is about.

5. If there is conflict present, it seems likely to have a good resolution. (If there is not conflict present, please choose "neutral.")

6. If this comment were automatically analyzed, it **should** be scored as a *negative* comment.

7. If this comment were automatically analyzed, it **should** be scored as a *positive* comment.

8. This comment is likely to be helpful to the person who wrote the proposal the comment is on.

We paid US$0.12 per comment reviewed. In this manipulation check, participants were allowed to participate up to 6 times, each time randomly selecting one of the six comment types not previously evaluated by that worker, and randomly selecting one of the four focal proposals. Participants in the manipulation check were not allowed to participate in the main experiment.

Manipulation Check on Comments: Results

160 unique Turkers, who passed the same inclusion restrictions applying to the main experiment, completed 432 rating tasks. Based in part on the results detailed below, we believe our comment type manipulations were successful.

Comparison between (focal) proposals

We checked each of the eight questions, plus time on task, and (using an ANOVA) looked for any significant differences between the four focal proposals. In cases where the Levene statistic rejects the null hypothesis that group variances are homogeneous, we use the Welch statistic as it is more appropriate than the *F* statistic to test for the equality of group means. We found only 3 significant differences (α=.05), which matches the expected number of randomly significant comparisons with 6 comment types and 9 comparisons each (6*9=54 comparisons). When all comment types were aggregated together, there were no significant differences between focal proposals on any of the 9 measures.

Author-directed vs. Content-directed: Questions 2 & 3

The author-directedness of the comment (Question 2) was significantly ($p<.0005$) higher in the author-directed conditions, considering the data overall (4.09 vs. 6.22 Likert points) or each proposal set independently, or each level of the "conflict" factor independently.

The content-directedness of the comment (Question 3) was significantly ($p<.0005$) higher in the content-directed conditions, considering the data overall (4.63 vs. 6.50 Likert points) or each proposal set independently, or each level of the "conflict" factor independently.

Conflict vs. non-conflict conditions: Questions 4 & 5

The comment's level of conflict (Question 4) was significantly ($p<.0005$) higher in the conflict conditions, considering the data overall (2.49 vs 5.53 Likert points) or each proposal set independently, or each level of the "direction" factor independently. The 95% confidence intervals for the mean of this item are fully on the "disagree" side for non-conflict conditions and on the "agree" side for conflict conditions.

Excluding non-conflict comment conditions from analysis, the conflict's likelihood of a good resolution (Question 5) was significantly ($p<.0005$) higher in the constructive conflict conditions than the "no good resolution" conditions, considering the data overall (3.16 vs 5.01 Likert points) or each proposal set independently, or each level of the "direction" factor independently ($p=.001$ in content-directed conditions).

Timing

The amount of time taken to complete the rating task was not significantly different across comment types, considering the data overall or each proposal set independently.

Comment type 4 more negative

The civility, positivity, and helpfulness of the comment (Questions 1, 7, & 8) were significantly ($p<.0005$) lower, and negativity (Question 6) significantly higher, in comment type 4 (author-directed unresolved conflict) than in the other comment types, considering the data overall or each proposal set independently.

Sentiment and Civility: Questions 1, 6, 7 & 8

All four of these questions were significantly correlated with each other ($p<.0005$ by Pearson or Spearman). Questions 6 and 7 were very strongly negatively correlated (Pearson's r: -.937; Spearman's ρ: -.930), as expected.

Comments were rated significantly ($p<.0005$) less civil, less positive, and more negative in conflict conditions than in non-conflict conditions, considering the data overall or each proposal set independently, or each level of the "direction" factor independently. Comments were rated significantly ($p<.0005$) less civil, less positive, and more negative in the conflict than non-conflict conditions even when excluding comment type 4 from analysis, overall or considering each level of the "direction" factor independently.

There are no civility or negativity differences between the two non-conflict conditions analyzing the data overall; the author-directed non-conflict condition is significantly ($p=.001$) more positive than the content-directed non-conflict condition.

In general, comments were scored as significantly ($p<.0005$) more civil and helpful when directed at the content than when directed at the author, overall and (with $p<.05$ on "civil") considering each proposal set independently.

There were no significant differences between author-directed and content-directed comments in positivity or negativity, considering the data overall.

The helpfulness of the comment (Question 8) was significantly ($.0005<p<.01$) lower in conflict than non-conflict conditions when analyzed overall.

MAIN EXPERIMENT RESULTS

After filtering as described above, we had 1252 responses. The median task completion time was 331.5 seconds, with an interquartile range of (246, 463) seconds.

Multi-Item Scales

In this subsection, we describe and validate the multi-item scales used in our primary dependent measure. We provide

evidence supporting the unidimensional treatment of our dependent construct "rated (or 'perceived') quality."

Cronbach's Alpha for the four-item feasibility scale is .869 (n=2391). An exploratory Principal Components analysis with these four items also showed all four loading onto a single principal component, and all four items were significantly ($p<.0005$) and strongly (average >0.6 by Pearson or Spearman) correlated with each other.

Cronbach's Alpha for the five-item quality scale, which includes the feasibility scale as one item, is .865 (n=2343). An exploratory Principal Components analysis with these five items also showed all five loading onto a single principal component, and all five items were significantly ($p<.0005$) and strongly (average >0.55 by Pearson or Spearman) correlated with each other. This is our multi-item perceived quality scale.

In order to increase comparability across the four proposal sets and reduce noise due to inherent differences in proposal quality, we compute the overall average quality rating for each proposal and compute each participant's rating as a deviation from that proposal-specific mean. This centered measure is our primary dependent variable for proposal quality evaluations discussed below. Unless otherwise noted, reported results are from planned contrast analyses within a larger ANOVA across comment types, with or without the assumption of homogeneous variances.

Main differences in ratings of focal proposal

Differences in the type of comment shown on the focal proposal caused significant differences in how participants rated the quality of that proposal. The conflict conditions led to significantly lower ratings of quality than the non-conflict or non-comment conditions (-.3071 vs .4334 Likert points, $p<.0005$), replicating the negative effect of conflict observed in [26] and **supporting H2**. (Table 2 below summarizes findings for each hypothesis.)

We did not observe any significant differences in the ratings of either proposal based on conflict resolution within conflict conditions, nor between "non conflict" and "no comment" conditions. Finding no significant difference in the latter comparison means we **do not find support for H1**.

Differences in ratings of second proposal

There was a small but significant ($p<.0005$, n=1249) correlation between participant's quality ratings of the first and second proposal ($r=.229$, $\rho=.194$ for deviation from proposal means as used here; $r=.188$, $\rho=.155$ for raw scale score), most likely indicating a per-rater effect that some raters were generally more generous and others more strict. Our experimental design controls for this by randomly assigning participants to experimental conditions, so we would not expect any systematic differences in rater scoring bias between the experimental conditions.

We did not find any statistically significant differences in the ratings of the second proposal caused by differences in the type of comment displayed on the first. Although differences in the rating of the first proposal based on comment type were observed, we did not find comment-caused differences in ratings of the second proposal even when the second proposal that was actually and was labeled as being by the same author or about similar topics. This is a **lack of support for H3, H4, and H5**.

Learning from comments

Most participants (>60% at each level of the "conflict" factor) agreed with the statement "I learned something from the comments." Participants reported strongest agreement (along the same 7-point Likert scale described above) with "I learned something from the comment(s)" in the "constructive conflict" comments, followed by the "non-conflict" and then "unresolved conflict" comments (group means of 5.18, 4.85, and 4.57 Likert points respectively, all differences statistically significant even with Tukey's Honestly Significant Differences test). This provides **support for H7.**

Further, participants' self-reported learning from the comments correlates significantly ($p<.0005$ overall and for each conflict type) with how much they thought "reading the comments *raised* my perception of proposal quality" ($r=.567$, $\rho=.550$, n=1044 overall). The correlation was strongest ($r=.768$, $\rho=.751$, n=333) in the non-conflict condition.

As expected, the degrees to which participants thought the comments *raised* and *lowered* their perception of proposal quality were significantly ($p<.0005$ in each conflict condition and overall, n=1053 overall) negatively correlated with each other ($r=-.427$; $\rho=-.449$ overall). The correlation was strongest in the constructive conflict condition ($r=-.689$; $\rho=-.697$, n=375). In general, significant differences on one or both of these measures were observed alongside significant differences in actual quality ratings, in a direction consistent with the observed effect, indicating **no support for H6**.

At least half the participants (at any level of the "conflict" factor) believed that the comments did not lower their perception of proposal quality (i.e. selected a "disagree" option on the "lowered" question) and most participants believed that reading the comments raised their perception of proposal quality. This aligns with the previous finding that participants believe seeing discussion increases their perception of the quality of the material discussed [26].

Power

Where we do not observe a difference between experimental groups, that logically means that either (A) no significant difference exists between the groups, (B) the experiment was not designed properly to detect a difference, or (C) the difference between groups was so small that our experiment did not have sufficient power to

detect it. (B) seems relatively unlikely given how closely our design mirrors a design that has previously identified differences [26] as well as the results of the manipulation check above. (C) appears relatively unlikely, based on computations using G*Power 3.1.9.2 [7] to identify how small of an effect we would have still had a 95% chance of observing, if it existed. For the ANOVA used here, the effect size parameter *f* is considered "small" at .1, "medium" at .25, and "large" at .40 [5:348].

Aggregating across relationships between proposals (Table 1 conditions) and looking for differences between second proposal ratings caused by comment types, we would have had a 95% chance of observing differences between cells with an effect size index of .13 or greater, and .10 or greater in a two-group comparison (e.g. conflict conditions vs. others). The minimum effect size observable with 95% probability is .17-.18 for a two-way comparison within any one of the Table 1 cells. The power for observing a medium size effect is 99.9% for a two-group comparison between comment types within any one of the Table 1 cells and higher for even a 7-group comparison aggregating across them.

Demographics

Prior work demonstrates that pre-existing cultural context can impact the way readers are affected by comments [25]. We report demographics here for easier comparison to other participant populations and more detailed identification of our participant pool. The initial questions and responses were as follows, except that questions 2 and 6 also each had an infrequently used "other" option manually recoded for analysis into the most similar available listed category.

1. Do you think that global warming is happening? {Yes: 85.7%, No: 6.5%, Don't know: 7.8%, n=1237}

2. Assuming global warming is happening, do you think it is … {Caused mostly by human activities: 49.8%, Caused mostly by natural changes in the environment: 10.3%, None of the above because global warming isn't happening: 2.7%, Caused by both human activities and natural changes: 37.1%, n=1250}

3. What is your gender? {Female: 50.4%, Male: 49.6%, n=1241}

4. What is your age? {under 20: 1.4%, 21- 29: 31.0%, 30- 39: 33.9%, 40- 49: 18.3%, 50- 64: 13.3%, 65 and over: 2.1%, n=1249}

5. What is the highest level of education you have attained to date? {High school or less: 17.1%, Attending college / university: 20.6%, Graduated from college / university: 46.3%, Attending graduate or professional school: 2.8%, Completed graduate or professional school: 13.3%, n=1245}

6. What is your current situation / status? {Student: 7.6%, Employed full-time: 54.7%, Employed part-time: 14.0%, Free-lance consultant: 7.5%, Unemployed [or homemaker]: 12.9%, Retired: 3.3%, n= 1244}

Last (after task). I am familiar with the MIT Climate CoLab. {Strongly disagree: 62.8%, Moderately disagree: 21.2%, Slightly disagree: 6.6%, Neutral 4.5%, Slightly agree: 2.9%, Moderately agree 1.1%, Strongly agree 0.9%, n= 1234}

Findings with Demographic Covariates

In addition to our primary hypotheses looking at how randomly assigned experimental conditions affect the target variables, we also did some analyses exploring how some of our observations may have interacted with demographic variables, which may help guide future research.

Among participants with no college degree, the constructive conflict led to significantly *lower* ratings than the unresolved conflict (contrast value .3865 points, p=.006-.008), while college graduates tended to rate the proposal quality significantly higher after viewing a non-conflict comment than after viewing no comment (contrast value .3141 points, p=.019-.021). Among participants reading comments illustrating unresolved conflict, college graduates rated the proposal slightly higher than non-college grads (contrast value .2950 points, p=.038), even though in the no-comment control condition college grads on average rated proposals slightly lower (contrast value .4044 points, p=.016-.022). This suggests that the more educated readers may have viewed the materials with a bit more critical and independent thinking.

Among comment conditions, the author-directed comments surprisingly led to higher quality evaluations than the content-directed comments in the "Constructive Conflict" conditions only, (-.4843 vs -.2048 Likert points, p=.017), which was enough to create an overall difference between content- and author-directed conditions (-.1564 vs. .0365 Likert points, p=.009). In further exploration, this significant difference in rating was only observed among female participants but not male participants, and among college graduates but not those without a college degree. Rater gender and college education were independent according to a chi-square test with a 95% chance of observing a "small" effect (effect size index w=.10 [7]).

Effect of Expertise

We found a small but statistically significant correlation between self-reported expertise on a proposal's topics and participants' quality rating of that proposal, both for the first proposal (r=.078 / p=.006; ρ=.093 / p = .001, n=1242) and the second proposal (r=.103 / p<.0005; ρ=.070 / p = .013, n=1250). There was also a small correlation between self-reported expertise on the first proposal's topics and if the comments raised participants' quality perceptions (r=.142; ρ=.133, p < .0005, n=1055). In general, more than three quarters of participants' self-reported topic-specific expertise level was on the "disagree" side of the scale and less than 14% of expertise self-reports were on the "agree"

side of the scale (n=2492). Especially combined with random assignment to the experimental conditions, we do not believe that pre-existing judgments about the material dominated quality evaluations (as when different materials were used in [25]).

CONCLUSIONS

#	Hypothesis
~~H1~~	Comment presence lowers 1st proposal qual. ratings
H2	Comment type & conflict affects size of H1 effect
~~H3~~	H1 effect extends to other proposals from platform
~~H4~~	Content-directed comments on one proposal affect quality ratings on a topically related proposal
~~H5~~	Author-directed comments on one proposal affect quality ratings on another proposal by that author
~~H6~~	Participants report perceptions being affected by comments in opposite direction from true effect
H7	Participants learn most from comments containing constructive conflict

Table 2: Summary of hypotheses. "No support" results where power calculations indicate a true effect would likely have been seen are indicated by ~~red strikethrough~~ on the left. Hypotheses supported by statistically significant differences are indicated by green underline in the left column.

In this large-scale [13] experiment, we showed that when comments about crowdsourced content are presented alongside that content, the contents of the comment affect how people perceive the content. We replicated a previous finding that comments containing conflict lower perceptions of content quality more strongly than comments which do not indicate conflict, but noted that readers reported learning more from constructive conflict comments than comments without conflict or without a good resolution. Participants in this study were also aware of the effects comments had on their perceptions of proposal quality, reporting significantly different answers (in the aligned direction) about how reading comments raised and/or lowered their perception of quality, when such differences were observed between their ratings.

We also observed that although the presence of comments may affect perceptions of the proposal the comment is on, that effect **does not** carry over to a second proposal read and judged in quick succession (in this case, on the same Web page), even when the second proposal is by the same author or about a similar mix of topics, and that connection between the proposals is called out with underlined bold header font. Participants in our experiment were apparently able to evaluate the second proposal independently after whatever effects the comment may have had on their perception of the first proposal. This result is encouraging for the future development of platforms that crowdsource proposals about how to solve some particular challenge, where readers may be evaluating those proposals either as part of a platform-hosted contest or for their own reasons (e.g. deciding what proposals to back or adopt in practice).

DISCUSSION

While large-scale observational data sets for many online interaction and content production platforms are now available, observational data does not allow us to distinguish between perceptions based on underlying qualities from those based on other factors which may affect perceptions [19:647]. Controlled experiments with random assignment of the factor being examined, as done here, allows us to draw causal conclusions about the differences between levels of the randomly assigned factor.

This study contributes to a small set of related studies in the field. For example, Steinfeld, Samuel-Azran, and Lev-On recently published what they described as the first study to examine how readers' perceptions of news articles' quality is affected by a set of comments presented below the article [25]. The study used eye trackers and post-study interviews to measure the attention users paid to the comments. Its setting intended to capture participants' common views about comments sections in online news sites, and found that most participants did not even read the comments, fewer than one in ten read any comments in detail, and even in those cases readers (undergraduates, mostly new first-years, at an Israeli college) often heavily discounted the content of the comments based on stereotypes about people who write them. As a result, that study found that the comments did not influence participants' evaluations of the articles, but drew a primary conclusion emphasizing the need to continue to "map the interplay between user comments and public opinion across various topics and domains" [25:72].

Past work at *CHI* specifically has looked at crowdsourced visual presentations of data and how comments on that data by other users affects later users' quantitative perceptions of graphical information. For example, Hullman, Adar, & Shah [13] asked Turkers to judge proportions or linear association strengths from charts, and found that these judgments were affected by a "social histogram" putatively showing prior viewers' estimates. However, the effects of biased information did not carry over to subsequent chart-perceiving tasks [13:1465]. That paper's Future Work section suggests exploring tasks more difficult than perception of quantitative information from the charts used there, where social influence is more likely to be observed. In this paper, we explored the task of evaluating the quality of a textually summarized proposal for addressing some aspect of a large, complex problem. Our results are consistent with that prior work about social influence on the perception of quantitative content [13], extending that result into more subjective judgments about content quality.

Social influence signals (e.g. comments) affect readers' perception of the content most directly related to the influence signals (e.g. proposals), as evidenced here. This experiment strengthens the call to be aware of these effects

and consider them when designing such platforms and the "need to form new theories and models that explain the impact of social processes on community-driven visualization environments and lead to new systems" [13:1469].

FUTURE WORK

Anchor points

This experiment found results for H1 and H6 that were different from our source for those hypotheses [26], despite a strong degree of similarity in task setup and associated details. When considering differences in task setup that may have produced these differences in results, the most likely source seems to be the difference in materials used, i.e. Climate CoLab proposals (and proposal comments) compared to Wikipedia articles (and Talk page discussions).

In contrast to participants in the present study based on Climate CoLab proposals, participants in the experiment based on Wikipedia materials were generally unaware of the effect, particularly the between-subjects primary observation that the presence of even non-conflict comments (in that study) led to lower perceptions of article quality [26]. The difference in H6 suggests that a different psychological process may have been operating with the different materials, and this may have produced the difference in H1.

For example, it may be that in the previous study [26], the presence of comments caused participants to become more acutely aware of Wikipedia articles' draft-in-progress status and anchor their initial quality measurements based on that category, while participants who saw no discussion evaluated quality beginning from an anchor point more applicable to a reference work perceived as polished or complete. In the present experiment, expectation-driven anchor points for the quality of a document described only as a "proposal," even without comments, may have already been lower and perhaps more in line with a draft or work-in-progress than a well-regarded reference work, so the mere presence of comments may not have lowered an initial anchor point. This could explain our finding of no support for H1. To determine if this indeed the case, future work would need to explore how anchor points are set for evaluation and how the way people perceive discussion may differ depending on the status of the artifact being discussed.

In a set of experiments, Galak and Nelson [10] asked readers to evaluate the quality of short stories, experimentally varying the fluency of the text through presentation factors like a more compressed font or asking participants to furrow their brow while reading. They find that the quality rating effects caused by differences in fluency vary as a function of what the reader expects from the reading and the anticipated purpose of the reading (e.g. conveying information contrasted with maximizing enjoyment). It is possible that different purposes for reading cause different expectations, and thus different anchor points from which perceptions are adjusted, between Wikipedia articles and Climate CoLab proposals.

Work in this area relates to the Affective Expectation Model [28], a theory about how people's expectations affects their subjective judgments. In this theory, people who have expectations about the quality value of some content (in the original paper, the humor value of comics) quickly check for features in given material that match their expectations and if found, perceive qualities based on those expectations, even inaccurately. This is a more specific version of schema theory applied to affect [28:524]. According to this theory, people making more specific evaluations (such as with the multidimensional rating scale readers evaluated on here) are more likely to notice discrepancies between their expectations and the actual material being evaluated [28:524], so the most surprising aspects of the theory are unlikely to be responsible for the differences observed here. However, a major part of this theory is about how *quickly* and *easily* people form judgments and how deeply their thoughts are engaged in the material [28:528] and measuring this requires more detailed control over the experimental environment (e.g. lab study) as opposed to Mechanical Turk where variance in e.g. response times can be attributed to a wide variety of other causes. Further research is necessary to better understand the expectation factors that could more clearly link these results into theories that explain the results based on reader expectations.

Larger Comment Sets

Because we used a randomized controlled experiment, we have some confidence the observed effects are *caused* by the presence and content of the comments. However, each participant saw only one comment posted on each proposal, while in real instances of the intended application environment there may be several comments posted on each proposal (as also seen in [25]), and readers may extract information from properties of the set, such as total/mean length, number of comments, unique participants, valence mean/variance, or other factors based on the interaction between commenters or between commenters and the proposal author, etc. This experiment held those factors relatively constant, and does not tell us if larger sets of comments with various properties would have led to different results. Future work would be needed to determine if more and/or more strongly critical comments have a cumulative effect.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from the National Science Foundation, grant #1302522. We also thank members of Daemo [9] for piloting the main experimental task, and the many Turkers who participated in either the main experiment or manipulation checks.

REFERENCES

1. Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy* 100, 5: 992–1026. Retrieved February 10, 2016 from http://www.jstor.org/stable/2138632
2. Dustin Carey. 2015. Comment 16: Language Requirement. Retrieved April 26, 2016 from http://climatecolab.org/contests/2016/shifting-behavior-for-a-changing-climate/c/proposal/1320702/tab/COMMENTS#_message_1349447
3. Climate CoLab. 2016. How will proposals be judged? Contest Rules. Retrieved March 2, 2016 from http://climatecolab.org/web/guest/resources/-/wiki/Main/contest+rules#Howwillproposalsbejudged
4. Climate CoLab. Climate CoLab Judges. Retrieved September 18, 2014 from http://climatecolab.org/resources/-/wiki/Main/Climate+CoLab+Judges
5. Jacob Cohen. 1969. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.
6. Erik P. Duhaime, Gary M. Olson, and Thomas W. Malone. 2015. *Broad Participation in Collective Problem Solving Can Influence Participants and Lead to Better Solutions: Evidence from the MIT Climate CoLab*. MIT Center for Collective Intelligence, Cambridge, MA. Retrieved from http://cci.mit.edu/working_papers_2012_2013/duhaime%20colab%20wp%206-2015%20final.pdf
7. Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods* 41, 4: 1149–1160. https://doi.org/10.3758/BRM.41.4.1149
8. David D. Van Fleet and Leanne Atwater. 1997. Gender Neutral Names: Don't Be So Sure! *Sex Roles* 37, 1–2: 111–123. https://doi.org/10.1023/A:1025696905342
9. S. Gaikwad, D. Morina, R. Nistala, M. Agarwal, A. Cossette, R. Bhanu, S. Savage, V. Narwal, K. Rajpal, J. Regino, A. Mithal, A. Ginzberg, A. Nath, K. R. Ziulkoski, T. Cossette, D. Gamage, A. Richmond-Fuller, R. Suzuki, J. Herrejon, K. V. Le, C. Flores-Saviaga, H. Thilakarathne, K. Gupta, W. Dai, A. Sastry, S. Goyal, T. Rajapakshe, N. Abolhassani, A. Xie, A. Reyes, S. Ingle, V. Jaramillo, M.D. Godinez, W. Angel, M. Godinez, C. Toxtli, J. Flores, A. Gupta, V. Sethia, D. Padilla, K. Milland, K. Setyadi, N. Wajirasena, M. Batagoda, R. Cruz, J. Damon, D. Nekkanti, T. Sarma, M.H. Saleh, G. Gongora-Svartzman, S. Bateni, G. Toledo-Barrera, A. Pena, R. Compton, D. Aariff, L. Palacios, M. P. Ritter, Nisha K.K., A. Kay, J. Uhrmeister, S. Nistala, M. Esfahani, E. Bakiu, C. Diemert, L. Matsumoto, M. Singh, V. Jaramillo-Lopez, K. Patel, R. Krishna, G. Kovacs, R. Vaish, and M. Bernstein. 2015. Daemo: a Self-Governed Crowdsourcing Marketplace. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (UIST '15). Retrieved May 12, 2016 from https://uist.acm.org/uist2015/schedule#uistp142
10. Jeff Galak and Leif D. Nelson. 2011. The Virtues of Opaque Prose: How Lay Beliefs About Fluency Influence Perceptions of Quality. *Journal of Experimental Social Psychology* 47, 1: 250–253. https://doi.org/10.1016/j.jesp.2010.08.002
11. David Hirshleifer and Siew Hong Teoh. 2003. Herd Behaviour and Cascading in Capital Markets: a Review and Synthesis. *European Financial Management* 9, 1: 25–66. https://doi.org/10.1111/1468-036X.00207
12. Daniel J. Howard and Roger A. Kerin. 2011. The effects of name similarity on message processing and persuasion. *Journal of Experimental Social Psychology* 47, 1: 63–71. https://doi.org/10.1016/j.jesp.2010.08.008
13. Jessica Hullman, Eytan Adar, and Priti Shah. 2011. The Impact of Social Information on Visual Judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 1461–1470. https://doi.org/10.1145/1978942.1979157
14. Joseph Kasof. 1993. Sex bias in the naming of stimulus persons. *Psychological Bulletin* 113, 1: 140–163. https://doi.org/10.1037/0033-2909.113.1.140
15. Aniket Kittur, Bongwon Suh, and Ed H. Chi. 2008. Can You Ever Trust a Wiki? In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work* (CSCW '08), 477. https://doi.org/10.1145/1460563.1460639
16. Thomas W. Malone, Robert Laubacher, and Laur Fisher. 2014. How Millions of People Can Help Solve Climate Change. *PBS NOVA Next*. Retrieved January 18, 2014 from http://www.pbs.org/wgbh/nova/next/earth/crowdsourcing-climate-change-solutions/
17. Thomas W. Malone, Jeffrey Nickerson, Robert Laubacher, Laur Fisher, Patrick de Boer, Yue Han, and W. Ben Towne. 2017. Putting the Pieces Back Together Again: Contest Webs for Large-Scale Problem Solving. In *Proceedings of the ACM 2017 Conference on Computer Supported Cooperative Work* (CSCW '17), In Press.
18. Pietro Michelucci and Janis L. Dickinson. 2016. The Power of Crowds. *Science* 351, 6268: 32–33. https://doi.org/10.1126/science.aad6499
19. Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social Influence Bias: A Randomized Experiment.

Science 341, 6146: 647–651. https://doi.org/10.1126/science.1240466

20. Pinar Ozturk, Yue Han, W. Ben Towne, and Jeffrey V. Nickerson. 2016. Topic Prevalence and Reuse in an Open Innovation Community. In *Collective Intelligence*. Retrieved from https://sites.google.com/a/stern.nyu.edu/collective-intelligence-conference/

21. M. J Salganik, P. S Dodds, and D. J Watts. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311, 5762: 854.

22. Clay Shirky. 2010. *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Penguin.

23. Jan-Benedict E. M. Steenkamp. 1990. Conceptual Model of the Quality Perception Process. *Journal of Business Research* 21, 4: 309–333. https://doi.org/10.1016/0148-2963(90)90019-A

24. Joel Stein. 2016. How Trolls Are Ruining the Internet. *Time*. Retrieved September 14, 2016 from http://time.com/4457110/internet-trolls/

25. Nili Steinfeld, Tal Samuel-Azran, and Azi Lev-On. 2016. User Comments and Public Opinion: Findings from an Eye-Tracking Experiment. *Computers in Human Behavior* 61: 63–72. https://doi.org/10.1016/j.chb.2016.03.004

26. W. Ben Towne, Aniket Kittur, Peter Kinnaird, and James Herbsleb. 2013. Your Process Is Showing: Controversy Management and Perceived Quality in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (CSCW '13), 1059–1068. https://doi.org/10.1145/2441776.2441896

27. W. Ben Towne, Carolyn P. Rosé, and James D. Herbsleb. 2016. Measuring Similarity Similarly: LDA and Human Perception. *ACM Transactions on Intelligent Systems and Technology* 8, 1: 7:1–7:28. https://doi.org/10.1145/2890510

28. Timothy D. Wilson, Douglas J. Lisle, Dolores Kraft, and Christopher G. Wetzel. 1989. Preferences as Expectation-Driven Inferences: Effects of Affective Expectations on Affective Experience. *Journal of Personality and Social Psychology* 56, 4: 519–530. https://doi.org/10.1037/0022-3514.56.4.519

29. Wolfram Alpha. 2016. Lee (given name). Retrieved April 28, 2016 from http://www.wolframalpha.com/input/?i=Lee+(given+name)

30. David L. Word, Charles D. Coleman, Robert Nunziata, and Robert Kominski. 2016. *Frequently Occurring Surnames Data*. US Census Bureau. Retrieved April 27, 2016 from http://www.census.gov/topics/population/genealogy/data.html

31. Gu Xu and Wei-Ying Ma. 2006. Building Implicit Links from Content for Forum Search. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '06), 300–307. https://doi.org/10.1145/1148170.1148224