Templates and Trust-o-meters: Towards a widely deployable indicator of trust in Wikipedia

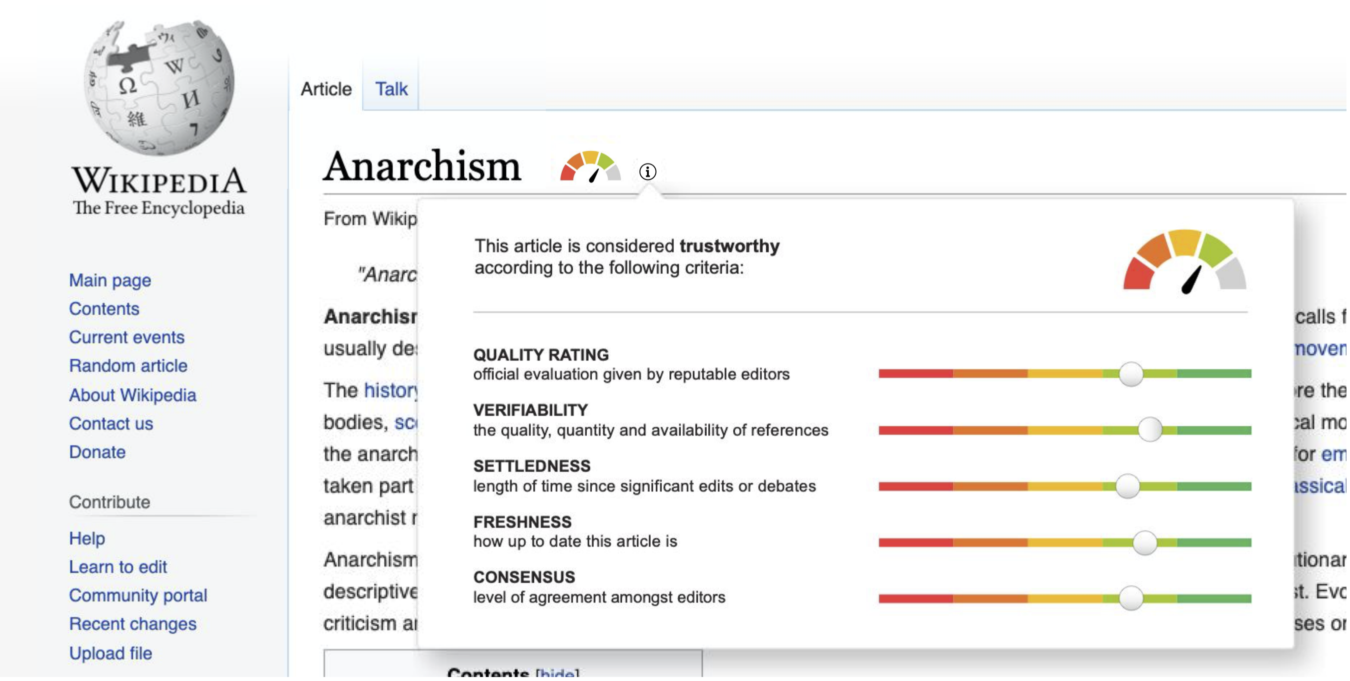
Andrew Kuznetsov
Carnegie Mellon University
Pittsburgh, PA, United States
adkuznet@cs.cmu.edu

Margeigh Novotny
Wikimedia Foundation
San Francisco, CA, United States
mnovotny@wikimedia.org

Jessica Klein
Wikimedia Foundation
San Francisco, CA, United States
jklein@wikimedia.org

Diego Saez-Trumper
Wikimedia Foundation
Barcelona, Spain
dsaez-trumper@acm.org

Aniket Kittur
Carnegie Mellon University
Pittsburgh, PA, United States
nkittur@cs.cmu.edu



This article is considered **trustworthy** according to the following criteria:

QUALITY RATING
official evaluation given by reputable editors

VERIFIABILITY
the quality, quantity and availability of references

SETTLEDNESS
length of time since significant edits or debates

FRESHNESS
how up to date this article is

CONSENSUS
level of agreement amongst editors

Figure 1: "Trust Gauge" with "Scoring Explanations" from Experiment 3, shown on Anarchism.

ABSTRACT

The success of Wikipedia and other user-generated content communities has been driven by the openness of recruiting volunteers globally, but this openness has also led to a persistent lack of trust in its content. Despite several attempts at developing trust indicators to help readers more quickly and accurately assess the quality of content, challenges remain for practical deployment to general consumers. In this work we identify and address three key challenges: empirically determining which metrics from prior and existing community approaches most impact reader trust; 2) validating indicator placements and designs that are both compact yet noticed by readers; and 3) demonstrating that such indicators can not only lower trust but also increase perceived trust in the system when appropriate. By addressing these, we aim to provide a foundation for future tools that can practically increase trust in user generated content and the sociotechnical systems that generate and maintain them.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517523

KEYWORDS

Wikipedia, User Trust, Interfaces

ACM Reference Format:

Andrew Kuznetsov, Margeigh Novotny, Jessica Klein, Diego Saez-Trumper, and Aniket Kittur. 2022. Templates and Trust-o-meters: Towards a widely deployable indicator of trust in Wikipedia. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3491102.3517523

1 INTRODUCTION

User generated content communities have driven the generation of tremendous value, including the largest encyclopedia in human history (Wikipedia) as well as thousands of other sources of knowledge that have become references for domains ranging from Star Wars [60] to neuroscience [75]. Their content benefits a variety of stakeholders both directly and indirectly, from general consumers [12] to artificial intelligence researchers [94]; from major internet sites [57][88] [30] to brands competing for recognition in the real-world [31].

Ironically, the same openness of recruiting volunteers from across the globe that has driven the success of user generated content communities can also be a perceived weakness. Since such communities involve volunteers with varied or unknown expertise, motives, and biases, a ubiquitous issue is a persistent lack of trust in their content [17] [42]. For example, despite the demonstrable success of Wikipedia, it suffers from a lack of trust from its own readers [20]. These effects are compounded by institutional issues such as students being advised to not trust content from Wikipedia [58]. The Wikimedia foundation has itself prioritized the development and deployment of trust indicators to address common misperceptions of trust by institutions and the general public in Wikipedia [62]. Such perceptions of mistrust are common for artifacts generated by other user generated content communities as well, ranging from Stackoverflow [87] and Yahoo Answers [36] to Youtube, Twitter [69], and Facebook [46].

There have been many attempts to address these challenges through the development of trust indicators that can help readers more quickly ascertain the accuracy and potential bias of user generated content. These efforts include measuring user activity [95]; the persistence of content [91]; content age [56]; the presence of conflict [10]; characteristics of the users generating the content [33]; content-based predictions of information quality [18]; and many more. With the continued development of new natural language processing and machine learning tools that can provide ever-increasing sophistication for inferring characteristics of content and the users who generate it (e.g., contextualized word embeddings of new content [71], active learning and statistical language models of revision histories [14], etc), and the ever-increasing sophistication of users who subvert such metrics to advance their own agendas (e.g., through sock puppets, bots, etc. [80] [47]), the iterative development of trust metrics is likely to continue to be an important area of research for the foreseeable future.

However, a key issue remains in translating these trust indicators from the lab into real world systems such as Wikipedia. While many of these trust indicators have been shown to have a significant impact on perceptions of trust, they have largely been in controlled studies where experimenters can use as much screen real estate as needed and participants are required to explicitly attend to and process the indicators, or are expert editors already invested in improving the system. Deploying such indicators in the wild raises three key 'last mile' problems that need to be addressed to make them viable and effective for general readers.

One key problem is deciding which information to provide to readers, who may have little time or attention for processing information outside what they were initially looking for. There have been several studies suggesting high level principles [17] and specific metrics [42] [3] [5] [4], for trust indicators, as well as as many notifications and banners that have been developed and deployed by the community. However, given the need to provide maximally impactful information with minimum time and attention costs, an important issue is determining which of these metrics to prioritize to the reader. In this work we draw on existing metrics from prior studies and commonly used Wikipedia templates, as well as a variety of potential new metrics, and empirically test their impact on perceptions of trust.

Another challenge lies in the tension between trust indicators taking up valuable screen real estate versus readers not noticing such information in the first place. Introducing a new visual element too large or intrusive can be distracting and take attention away from or push the content that both creators and readers want to focus on 'below the fold'. On the other hand, readers are likely to ignore new visual elements that are not sufficiently salient or that don't appear to provide valuable information at a glance ("banner blindness") [7]. Developing such a compact yet salient form (which includes what information elements should be included, what visual form they should take, what interactions they afford, and where they should be placed on the page) remains an open challenge; previous indicators have instead largely focused on evaluating their value and accuracy while relying on attention derived from either experimenter control or the intrinsic interest of expert editors. However, addressing this challenge may be of fundamental importance to widespread deployment for general readers.

Finally, for a community to choose to deploy such indicators it is important to have evidence that such indicators can increase perceptions of trust in appropriate articles or the system as a whole, rather than only decreasing trust. One core challenge here is the general concern that exposing internal processes about how content is created can be undesirable. Indeed, Jimmy Wales, a founder of Wikipedia, stated that Wikipedia is "like a sausage: you might like the taste of it, but you don't necessarily want to see how it's made" [89]. While significant prior work has shown that surfacing negative issues about articles' internal processes can reduce perceptions of trust in those articles, evidence on whether the positive version of such information could also increase trust is both sparse and mixed. For example, when readers come to realize that Wikipedia is not written by professionals but by others like themselves it can erode their trust in the content [85]. In other words, the very lifeblood of Wikipedia – the lively debate and discussion and vetting that enables it to fairly process potentially biased information at scale – may result in a self-defeating perception of untrustworthiness. This uncertainty and potential negative bias suggests a fundamental issue to the deployment of any trust indicator: system designers are unlikely to deploy such indicators if they are only likely to reduce trust for content that has issues, but not increase trust in more reliable content. Thus, demonstrating a trust indicator can both increase and decrease trust appropriate to different quality content is a vital step forward. Here we introduce and validate one example of a trust indicator that has these properties, characterizing its response across its range of values and across varied content.

In summary, in this paper we explore the above challenges for the 'last mile' problem of making trust indicators for user generated content impactful and deployable to general consumers. We focus on the specific context of Wikipedia, a user-generated content community that has produced millions of encyclopedic articles from the efforts of millions of volunteers across the globe. Wikipedia has a number of characteristics that make it a suitable platform for this study. It is widely accessed, with billions of page views per year [77] [1], and thus provides environmental validity and a large pool of potential readers for empirical testing and iterative design. It also has a sophisticated internal system of trust indicators consisting of notifications, banners, and assessments aimed at providing users with accurate information about the quality of its content. Although many of those existing indicators are aimed at internal editors rather than general readers, they represent information that might inform a new trust indicator for general readers. The lessons learned about Wikipedia may generalize to other wiki systems, which often use similar software, as well as to user-generated content communities more generally which employ volunteers to create artifacts of lasting value ranging from open source software to question-answer forums. Understanding and addressing these challenges can thus help provide a foundation for future approaches that aim to practically increase trust in user generated content and the systems that generate and maintain them. Our contributions in this paper include:

- Empirically comparing the impact of surfacing both existing and potential signals of article quality on users' perceptions of trust
- Developing and validating a compact yet salient trust indicator
- Demonstrating an indicator's impact on readers in both positively as well as negatively impacting perceptions of trust, and characterizing readers' response curves across a variety of values, content, and prior experience.

2 RELATED WORK

2.1 Determining the Credibility of Digital Information

Despite its relative ease of access, digital information is at a disadvantage as readers perceive electronic contexts to be less trustworthy [78]. As a result, trust is an integral dimension of any piece of online information - researchers widely cite it as a mediating variable between information quality and use [40], often highlight its significant financial consequences [48], and actively explore the underlying factors that influence trust in online information. Fogg et al. identified a number of salient website features related to credibility [21]. Several frameworks have since emerged (e.g. [86], [73]) with many specifically created to assist users in creating more trustworthy content (e.g. [84]). Fogg et al.'s well-cited definition of computer credibility defines it as a combination of two factors: trustworthiness and expertise [22]. However, many of the world's websites now primarily rely on their userbases for content generation - whose expertise (and identity) is often poorly established or unverifiable.

2.2 Trust Research on Wikipedia

Despite the participatory nature of peer production, websites that host collaborative user-generated content have created tremendous value from open source software [6], to 3D printing [61] [43], and more than 12,000 knowledge base 'wikis' in a variety of domains and languages [79]. The largest and most prominent of such wikis and website communities is that of English Wikipedia, with approximately 500 million edits and about 200 billion pageviews yearly [1]. As the world's largest collaborative user-generated content community and knowledge base, Wikipedia's trustworthiness has been the subject of a significant amount of research. For this reason and for reasons of scope, we focus our study of trust in the artifacts of such communities to that of Wikipedia. The site has been compared by researchers to Encyclopedia Britannica in terms of accuracy [25], perceived credibility [20], and bias [27].

Researchers have attempted to understand components of the Wikipedia experience most influential in user trust. Denning et al. aggregated a number of risks with Wikipedia use, with many involving credibility assessment [17]. Lucassen et al. identified salient Wikipedia article features related to reader trust such as images, references, and textual features, with think-aloud protocols and surveys in a lab study with university students [53]. Some researchers have examined Wikipedia's social aspect. Jessen et al. suggest Wikipedia's trustworthiness comes from many small indications of social validation [37](e.g. a large set of editors are unlikely to be conspiring to insert the same incorrect fact), consistent with recent literature suggesting information trust assessments are a social activity, often reliant on cues from others [59]. Towne et al. demonstrated that surfacing editor discussions from Wikipedia talk page content significantly sways reader's perceptions of article trustworthiness [85].

Our work builds on the research above to identify signals from past approaches that may be useful to include in trust relevant indicators, and goes beyond it to empirically investigate the relative impact of these different signals on perceptions of reader trust. Furthermore, we also investigate the relative impact of existing templates used by the Wikipedia community to indicate potential quality issues, which many readers are unaware of [58], as well as several types of plausible notices not currently implemented by the editor community.

2.3 Automated Wikipedia Trustworthiness Assessments

A number of researchers have explored the creation of computer algorithms to automate the assessment process and surface insights. Zeng et al. suggested that low-level revision metrics can accurately detect trustworthiness changes of an article, although this was only validated by the author's own eyes [95]. Similarly, Kramer et al. calculated a set of trust metrics from phrasal analysis of article revision history but only validated by inspecting 3D visualizations

[1] https://stats.wikimedia.org/#/all-projects

of the metrics computed for four example articles [44]. Researchers have more deeply validated this approach by calculating singular trustworthiness metrics; Dondio et al. were able to successfully differentiate high and low quality articles using low-level statistical features over 8,000 articles representing 65% of overall editing activity[18]. Although text features are popular, others have also centered their approach on authors and references. Hu et al., examined calculating article quality by way of computing author credibility with edit histories [33]. Moturu et al. explored a dispersion degree score (DDS) model that uses approaches using source and citations quality (e.g. density of citations per paragraph) and author credibility (e.g. proportion of unregistered editors with a single edit) [65]. More recent work by Dang et al. has demonstrated statistical approaches that do not require identifying and crafting specific textual features [16]. All of these systems have paved the way for official systems sponsored by Wikipedia and available for use by editors, such as Wikipedia's ORES model [28].

However, there remains a 'last mile' issue with practically deploying such metrics into systems that are effective for general readers. For example, while Adler et al.'s WikiTrust system combined an accurate editor consensus algorithm with an interface that centered on the use of orange highlighting [2] [3]. However, eye-tracking research by Lucassen et al. later found that while pages with more orange on the page were rated as less trustworthy, readers assessed the tool's utility as low and reported a lack of clarity in their perceptions [54]. Our work aims to identify and address factors relevant to these last mile issues with trust indicators such that users notice them, find useful and easily comprehensible information, and can more accurately calibrate their trust perceptions to the quality of the page, both positively and negatively.

2.4 Wikipedia Trust Support Interfaces

To better understand the 'last mile' of an end-to-end system designed to assist with trustworthiness assessments, researchers have studied a number of approaches for surfacing relevant page-related calculations to readers. Kittur et al. successfully explored the use of a large static visual element that displayed a number of aggregated trust-relevant metrics in high detail, finding the element to influence perceived trustworthiness in both positive and negative directions [42]. Other researchers have explored the use of a large, multi-metric visualization panel on the left side of the article [13] for rapid trustworthiness judgements. Researchers have also explored the use of separate interactive dashboards that surface editor activity, such as Wikidashboard [83], later shown to successfully increase article and author credibility judgements [74]. Contropedia [10][9][8], another interactive dashboard, focused on surfacing terms and article content that was often-mentioned in discussions and editing activity. However, the large size and complexity of these elements and dashboards limits their suitability for communicating metrics to a general readership audience as part of a widespread deployment, the design goal of this paper.

3 EXPERIMENT 1

In experiment 1, we aim to identify the relative impact of surfacing various trust-relevant metrics on perceived user trust, using visual 'single-issue' notices. Attempts to help users better understand the accuracy and trustworthiness of user generated content in Wikipedia span a wide variety of approaches. However, comparing the relative impact of dozens of trust metrics and thousands of existing Wikipedia 'banner template' notices was infeasible for the scope of this paper. Instead, to represent current practices we measure and compare to existing issue-related notices that are popular across Wikipedia. We also generate a set of 'single-issue' notices aimed at probing the space of potential future measures (e.g., trust metrics from prior or plausible future work) one at a time.

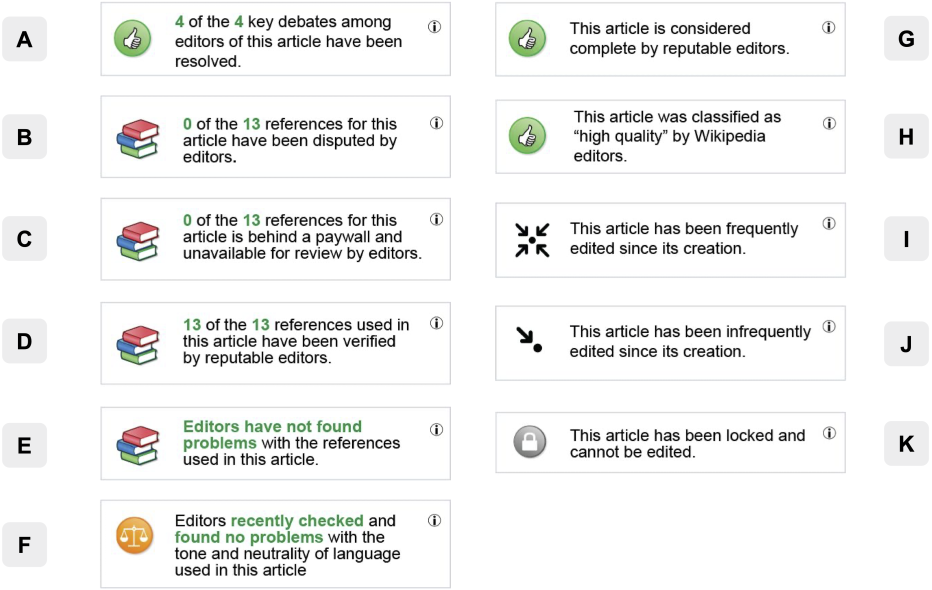
To represent existing notices commonly used by editors, we exhaustively collected all issue-related templates from the Wikipedia Template Index [2], curating 307 unique templates in total. Using a transclusion counter [3], we identified the top 20 most commonly embedded templates. This selection was further refined based on the following criteria: (1) the template must highlight an article issue, (2) it must focus on one specific issue (3) it must not only be relevant to a specific article type, (4) it must be intended for article use, not for sections or lists. This procedure generated nine commonly-used templates that are included on over 480,000 pages as of August 2020. They are shown in Figure 3. We include the existing 'multiple issues' template as the sole exception to the second rule above as it does not explicitly reference the multiple issues implied and could help estimate the impact of surfacing multiple issues at once without the potentially complex interactions between different specific issues. The popular 'unreferenced' (no references) template was excluded as it would not be believable without modifying the article references. The article banners used in this experiment were the most basic versions available; editors can sometimes insert additional details into each using a parameter.

To probe additional parts of the design space we also designed a set of 'single issue' notices similar in form and focus to the existing Wikipedia Templates but containing additional information we believed had the potential to impact trust both positively and negatively, depending on the information contained therein. These notices were generated through an iterative design process informed by converging evidence from a variety of sources that included:

- prior work on key risks perceived by users in wiki systems, such as questions about source validity and editor reliability (e.g. [17]);
- in-depth coding of several dozen discussion pages for key issues and concerns;
- analysis of several thousand edit history comments for the same pages, which often contain key issues such as reasons for reverting questionable references;
- and Wikipedia community policy documents including the Wikipedia Guidelines for Creation and Deletion, Wikipedia's Five Pillars (a community values document), as well as the general article quality assessment and promotion guidelines.

From these we focused on elements where a potentially viable algorithmic or sociotechnical path to generating the data for that signal at scale could be proposed. From this process we derived 10 article-related pieces of information on a diverse set of information from reference issues to resolution of key debates. For each new

[2] wikipedia.org/wiki/Wikipedia:Template_index
[3] templatecount.toolforge.org



4 of the 4 key debates among editors of this article have been resolved.

0 of the 13 references for this article have been disputed by editors.

0 of the 13 references for this article is behind a paywall and unavailable for review by editors.

13 of the 13 references used in this article have been verified by reputable editors.

Editors have not found problems with the references used in this article.

Editors recently checked and found no problems with the tone and neutrality of language used in this article

This article is considered complete by reputable editors.

This article was classified as "high quality" by Wikipedia editors.

This article has been frequently edited since its creation.

This article has been infrequently edited since its creation.

This article has been locked and cannot be edited.

Figure 2: Notices generated in the positive intervention condition of Experiment 1. Notices for the negative condition were generated by altering relevant values, for example changing 0 of 13 references disputed to 7 of 13.

notice, a 'high trust' and 'low trust' version was created by averaging the relevant manual calculations between articles of high and low quality (respectively). Our notices are shown in Figure 2, and include information about the presence of: 1) editor debates and disputes, 2) issues with the article and its references, 3) assessed quality or completeness by editors, and 4) temporal information about recent or present issues or editing.

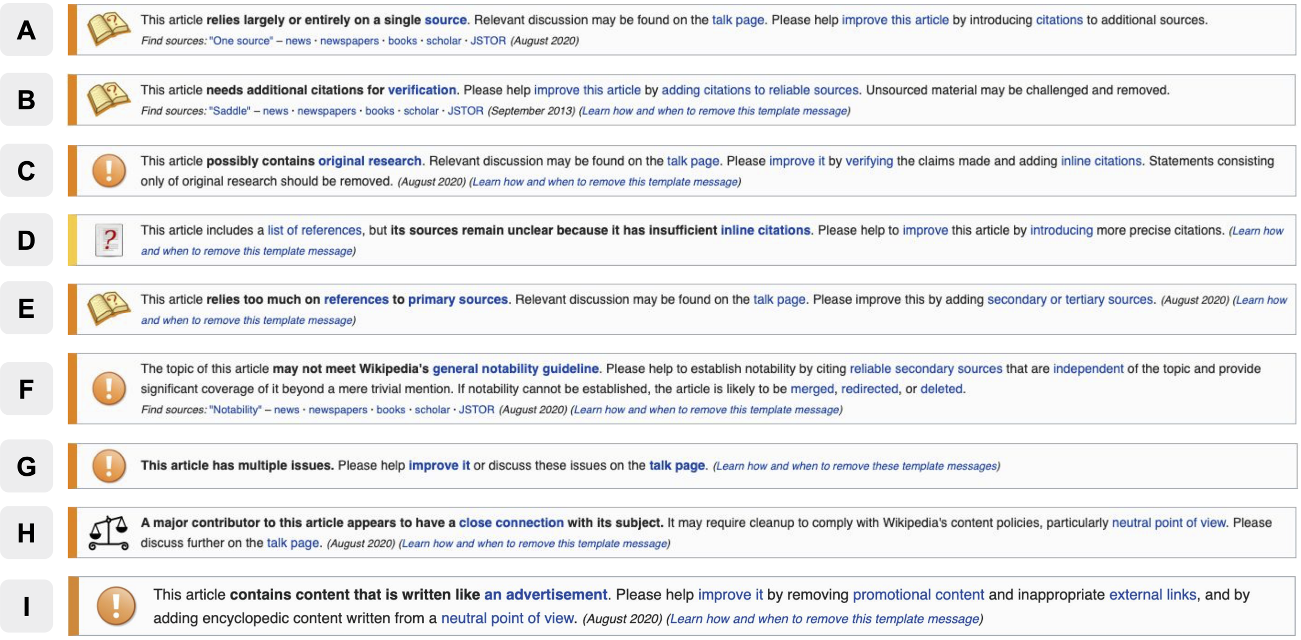
The newly-created notices are different from the article banner templates in a number of ways. Article banners are primarily designed to motivate, recruit and assist editors in fixing major issues on tagged articles. As a result they are visually complex and occupy a large footprint in the most visible section of the page. In comparison, the notices we created are designed for frequent reader usage: they are comparatively much smaller, visually neutral, and visually simplified to a single icon with a clear, color-coded statement that avoids terminology. Additional context in the notices is designed as an informative and granular metric for reader judgement, not editor motivation. Since the information does not need to inspire a high-effort action (editing), the notices logically support a much larger range of metrics and values (e.g. positive, reassuring information that celebrates or contextualizes the article as compared to only surfacing concerns that must be resolved).

We collected reader trustworthiness ratings for a set of pages in which we added interventions, either from our generated set or from existing Wikipedia notices. To focus reader attention on the interventions, all existing issue-related or article status templates were removed (e.g. edit locks and featured stars). In addition, all hyperlinks in the article (including those to the edit history and talk page) were disabled to isolate the notices as sources of information. The content of the article itself was unmodified between conditions, only the notices at the top were varied. Here we were interested in the impact on trust of a notice given that the user had seen the notice; we consider the salience of the notice in Experiment 2. To address this, in conditions where an intervention was present, participants were asked two verifiable questions requiring them to recall information present in the notices, with the second typically requiring the recall of a specific statistic present in the notice. Furthermore, to ensure that participants had viewed the article, the survey asked readers several verifiable questions related to the content of the article ('how many sections / images / references were present in the article') similar to those asked in [42]. Following these questions, readers were asked for their perception of the article's trustworthiness on a 7-point scale. To further verify that the intervention was salient, participants were also asked to explain their trustworthiness rating. Both of these measures are adapted from prior work [42], enabling us to collect trust judgements from participants with minimal survey fatigue. Specifically, the survey asked for a response to the question "I believe this article is trustworthy" on a 7-point scale (from "Strongly Agree" to "Strongly Disagree") and then asked participants to elaborate in the latter question of "Please explain your answer to the previous question."

3.1 Design

The articles shown to participants were chosen from a set of four in a 2x2 design, with each participant seeing an article that was either high or low quality and either high or low controversy. Each reader was shown one article and either one or no interventions (baseline).

To limit variance and to be consistent with prior work on Wikipedia trust visualization [42], one article was selected for each combination of high and low quality and controversy. High quality articles were randomly sampled from a list of articles assessed at 'GA-class'



A This article **relies largely or entirely on a single source**. Relevant discussion may be found on the talk page. Please help improve this article by introducing citations to additional sources.
Find sources: "One source" – news · newspapers · books · scholar · JSTOR *(August 2020)*

B This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Saddle" – news · newspapers · books · scholar · JSTOR *(September 2013) (Learn how and when to remove this template message)*

C This article **possibly contains original research**. Relevant discussion may be found on the talk page. Please improve it by verifying the claims made and adding inline citations. Statements consisting only of original research should be removed. *(August 2020) (Learn how and when to remove this template message)*

D This article includes a list of references, but **its sources remain unclear because it has insufficient inline citations**. Please help to improve this article by introducing more precise citations. *(Learn how and when to remove this template message)*

E This article **relies too much on references to primary sources**. Relevant discussion may be found on the talk page. Please improve this by adding secondary or tertiary sources. *(August 2020) (Learn how and when to remove this template message)*

F The topic of this article **may not meet Wikipedia's general notability guideline**. Please help to establish notability by citing reliable secondary sources that are independent of the topic and provide significant coverage of it beyond a mere trivial mention. If notability cannot be established, the article is likely to be merged, redirected, or deleted.
Find sources: "Notability" – news · newspapers · books · scholar · JSTOR *(August 2020) (Learn how and when to remove this template message)*

G **This article has multiple issues.** Please help **improve it** or discuss these issues on the **talk page**. *(Learn how and when to remove these template messages)*

H **A major contributor to this article appears to have a close connection with its subject.** It may require cleanup to comply with Wikipedia's content policies, particularly neutral point of view. Please discuss further on the talk page. *(August 2020) (Learn how and when to remove this template message)*

I This article **contains content that is written like an advertisement**. Please help improve it by removing promotional content and inappropriate external links, and by adding encyclopedic content written from a neutral point of view. *(August 2020) (Learn how and when to remove this template message)*

Figure 3: Existing Wikipedia article templates tested in Experiment 1. All templates reflect problematic issues with an article ranging from relying on a single source to being written like an advertisement.

[4], while low quality articles were randomly selected from a list of 'C-class' articles. To determine controversial articles, a pretest was conducted using a separate group of participants from Amazon's Mechanical Turk using a questionnaire that asked participants to rate the subject of the article (not the article itself) for Controversy and Expected Article Trustworthiness (n=20 participants). Article topics with high controversy and low expected trustworthiness were selected for the controversial dimensions to maximize the potential impact of intervention notices. Articles were selected such that the entire set were similar in length and contained a similar number of images in order to equate surface signals of article credibility [50], particularly for novices [51]. This cycle was repeated until a suitable set of articles was selected. The articles selected are listed in Table 1. The average perceived controversy scores of each was as follows (higher is more controversial): -2.67 (World's littlest skyscraper), -0.89 (Spotted Saddle Horse), 1.22 (Fursuit), and 1.65 (Discrimination against asexual people). To minimize variance between notices within conditions, a single 'High Trust' and 'Low Trust' visualization of each metric was created by calculating average percentage metrics for the high and low quality groups and applying this average. For example, 44.7 percent of articles were behind a paywall for the two high quality articles, and thus displayed as 17/38 articles on the skyscraper article. The visualizations that would be present on a high quality article can be viewed in Figure 2. All interventions were placed in the standard template location at the top of the article. The nine Wikipedia template interventions were classified as 'low trust' visualizations as all highlight negative issues. These can be viewed in Figure 3.

3.2 Participants

Participants were recruited through Amazon's Mechanical Turk (MTurk) web service. Inclusion was restricted to US-based users with more than 10,000 approved tasks and above 97% task approval who had not previously participated in this experiment. A total of 1,491 users participated and were compensated at $15 an hour as estimated by pretesting the time it took to complete the task with a member of the research team. Information available via TurkerView[5], a website that crowdsources MTurk task hourly wages with a browser extension, suggests that the hourly rate estimate was met, and likely significantly exceeded. Each participant only evaluated one article and intervention combination.

3.3 Results

3.3.1 Attention checks. We noticed during the experiment that, despite multiple rounds of iterating on task design and attention checks, a surprisingly large proportion of participants still did not seem to have attended to the notices. To better characterize this, in addition to the attention check questions a member of the research team and an external coder examined the rating explanations of workers to check for explicit references to the intervention. This was done in one coding round followed by a reconciliation phase to resolve any differences. The percentage of readers who had not seen the intervention completely was 48.5%. We found this

[4] 'GA-class' articles are assessed by Wikipedia editors to be of high quality

[5] https://turkerview.com/

	High Quality	**Low Quality**
Controversial	Fursuit	Discrimination against asexual people
Uncontroversial	Spotted Saddle Horse	World's littlest skyscraper

Table 1: Articles used in Experiment 1 interventions, sampled from differing levels of Wikipedia quality and perceived controversy.

surprising, as our notices (including existing Wikipedia templates) were placed in a high visibility location where current Wikipedia templates reside and multiple task design elements were put in place to help participants focus on them. In the below sections, all results are reported for only those participants that passed the two attention checks described above. Readers not noticing an indicator is a challenge that we return to and address directly in Experiment 2.

3.3.2 Interventions. Interventions were compared to a baseline measurement (collected when no notice or banner is shown on the article) using Dunnett's method to adjust for multiple comparisons. The existing Wikipedia banners, the new positive 'single-issue' notices, and the new negative 'single-issue' notices were all separately compared to the control condition using this method. As expected, several of the existing Wikipedia templates significantly influenced reader trust in the negative direction. This is unsurprising, as these templates are designed to indicate a serious issue and inspire editors to mobilize. The remaining templates, 'Additional citations', 'Inline citations', 'Notability', 'Original Research', 'Too Reliant on Primary Sources' and 'Too Reliant on Single Source' did not result in significant changes. It is possible that the specific terms used in these templates were confusing to the casual readers taking the survey. Particularly strong effects were noted in 'Multiple Issues' (-2.101; 'Moderately Lowered', p<0.001), 'Written like Advertisement' (-1.937, p<0.001), and 'Conflict of Interest' (-1.182, p<0.05).[6] It is possible these issues form a set of more serious 'grave' issues that readers consider to be particularly problematic, in particular core violations of the article's good-faith in aggregating information for the user.

There were also many newly designed negative notices that produced significant negative effects. The strongest negative effects were found in 'Editor Disputed References' (-1.601 points from baseline, p<0.001), 'General Reference Issues' (-1.444, p=0.002), 'Tone and Neutrality Issues' (-1.184, p=0.012), and 'Assessed as Complete' (-1.101, p=0.017). Participants appeared to understand the notices and were concerned by them:

> "I think that 13 out of 38 is not a good ratio. I am suspicious of the accuracy of the article and would seek out more information to confirm the truth."
>
> "If a third of the references were disputed I'd have to wonder about the credibility of the article."

Surprisingly, no individual positive notices resulted in significant changes in perceived trust, including notices detailing the absence of issues that when present did cause significant negative changes in trust. There are several possible explanations for this finding. It may be that individual notices that bring up a potential issue, even if reassuring about that specific issue, open the door for participants to think about other issues that might be problematic for this article that were not surfaced. Even if no specific other issues are made salient, individual notices could prime more general concerns about Wikipedia's editing environment:

> "I believe it is trustworthy because it has been checked but I do know wikipedia can be edited by anyone so it wont always be 100 percent trustworthy"

It is also possible that we did not have enough power to detect a significant statistical difference either due to insufficent participants numbers or the design of the survey instrument. In analyzing participants' responses some seemed to make positive inferences about the process by which the article was generated:

> "I believe it is trustworthy because there are no problems found with this article, as per the button on top."

An alternative explanation is that some individual notices surfaced aspects of Wikipedia's process that might be unfamiliar to readers, such as the presence of debates or vetting of references. Without additional context about the importance of such signals, such as positioning them with respect to the settledness or verifiability of the article, general readers might not know how to interpret these notices. Finally, in this experiment we asked only about participants' perceptions of the specific articles that the notices were applied to. It is possible that, even if not increasing trust in a specific article, the presence of a positive notice might engender global increases in trust in Wikipedia as a whole. Such an argument is possible to make even for negative notices if only a small proportion of articles have negative notices and their local negative impact is outweighed by a global increase in trust overall. In combination with the above considerations, in Experiment 3 we more directly test these assumptions through exploring an aggregate indicator that might provide a stronger intervention and impact trust positively as well as negatively.

4 EXPERIMENT 2: PLACEMENT AND INTEGRATION

As described above, one significant challenge we encountered in Experiment 1 was that despite multiple attention checks nearly 50% of readers had not noticed the intervention. Historically, this effect of "banner blindness" is a longstanding UX issue for visual messaging [7] and represents another significant 'last mile' challenge for any potential trust indicator. In Experiment 2, we address this by investigating factors relating to the placement and form of a trust indicator, in particular aiming to probe the trade-off between its salience and compactness. Such a trade-off is critical in enabling viable deployment, aiming to get the best of both worlds in terms of not being distracting or interfering with valuable content, but also being noticeable to general readers.

[6] Significance judgments for all analyses were computed with Dunnett's Method to adjust for multiple comparisons through the R package 'multcomp'.

Wikipedia Banners

Name	N	Mean Baseline Difference	Standard Error	P-Value
Baseline (No notice or banner shown on article)	43	-	0.213	-
[G] Multiple Issues	**14**	**-2.101**	**0.449**	**<0.001**
[I] Written Like Advertisement	**26**	**-1.937**	**0.363**	**<0.001**
[H] Conflict of Interest	**16**	**-1.182**	**0.428**	**0.050**
[A] Too Reliant on Single Source	18	-1.078	0.410	0.072
[F] Notability	26	-0.821	0.363	0.176
[B] Additional Citations	25	-0.664	0.367	0.427
[E] Too Reliant on Primary Sources	24	-0.536	0.372	0.704
[C] Original Research	12	-0.411	0.477	0.978
[D] Insufficient Inline Citations	20	-0.394	0.395	0.947

Table 2: Effects of existing article banner templates in Experiment 1. Significant effects are bolded. Each banner can be found by matching the preceding alphabetic tag with the respective banner in Figure 3.

Notices (Negative Versions)

Name	N	Mean Baseline Difference	Standard Error	P-Value
Baseline (No notice or banner shown on article)	43	-	0.213	-
[B] Editor Disputed References	**21**	**-1.601**	**0.381**	**<0.001**
[E] General Reference Issues	**20**	**-1.444**	**0.387**	**0.002**
[F] Tone and Neutrality Issues	**25**	**-1.184**	**0.360**	**0.012**
[G] Assessed as Incomplete/Complete	**18**	**-1.101**	**0.347**	**0.017**
[A] Resolved Key Debates (Low/High)	30	-0.811	0.340	0.152
[I] Frequently Edited	21	-0.792	0.381	0.293
[H] Needs Improvement/High Quality	31	-0.712	0.337	0.274
[K] Edit Lock	15	-0.211	0.429	1.000
[D] References Verified By Reputable Editors (Low/High)	14	-0.173	0.440	1.000
[C] Unavailable References (Low/High)	14	-0.173	0.440	1.000
[J] Infrequently Edited	29	0.049	0.344	1.000

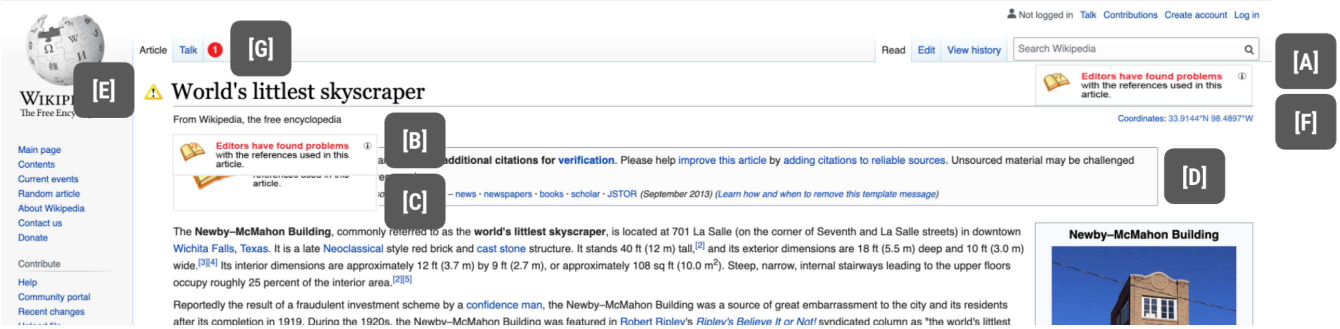
Table 3: Effects of negative newly-created notices in Experiment 1. Significant effects are bolded. The positive versions of each notices can be found by matching the preceding alphabetic tag with the respective notice in Figure 2.

The notices and banners in the previous experiment were placed at the top of the article, a common practice throughout Wikipedia for issue-related templates. However, notices in this space can be distracting, take up valuable 'above-the-fold' screen space, and frustrate editors if large and bold[7]. To better understand how an indicator could be made visually salient with a low footprint, we tested the salience of indicators in various 'high-footprint' and 'low-footprint' forms inspired by existing Wikipedia notifications and alerts:

- As a notice in the 'high-footprint' standard location of Wikipedia issue templates
- As a notice in a 'low-footprint' placement tucked into the right-hand side of the page
- As a compact triangular warning icon used by Wikipedia maintenance templates prominently in front of the article title. Interacting with this icon would activate a notice.
- As a compact but brightly-colored social media style notification badge placed next to the Article and Talk page tabs. Interacting with this icon would activate a notice.

We augmented this set with a version of the standard location notice that was twice as tall to probe for any size-related changes to salience, as well as a version of the right-hand notice that slowly blinks to test the effect of motion and to validate users' ability to notice content in the 'low-footprint' placement for the notice. Together with the standard Wikipedia issue template as a point of comparison, we tested seven placements in total, summarized in Table 4 with Figure 4 demonstrating all placement positions overlaid.

[7] Related unwelcome activity is sometimes referred to by Wikipedians as 'Tag Bombing' or 'Drive-by Tagging' with the offending party branded a 'WikiImp'.



World's littlest skyscraper

From Wikipedia, the free encyclopedia

Editors have found problems with the references used in this article.

additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged

– news · newspapers · books · scholar · JSTOR (September 2013) (Learn how and when to remove this template message)

The **Newby–McMahon Building**, commonly referred to as the **world's littlest skyscraper**, is located at 701 La Salle (on the corner of Seventh and La Salle streets) in downtown Wichita Falls, Texas. It is a late Neoclassical style red brick and cast stone structure. It stands 40 ft (12 m) tall,[2] and its exterior dimensions are 18 ft (5.5 m) deep and 10 ft (3.0 m) wide.[3][4] Its interior dimensions are approximately 12 ft (3.7 m) by 9 ft (2.7 m), or approximately 108 sq ft (10.0 m^2). Steep, narrow, internal stairways leading to the upper floors occupy roughly 25 percent of the interior area.[2][5]

Reportedly the result of a fraudulent investment scheme by a confidence man, the Newby–McMahon Building was a source of great embarrassment to the city and its residents after its completion in 1919. During the 1920s, the Newby–McMahon Building was featured in Robert Ripley's *Ripley's Believe It or Not!* syndicated column as "the world's littlest

Figure 4: Placements tested in Experiment 2 as per Table 4 (placements are overlaid for illustration).

The different placements all included a notice regarding general reference issues - selected due to its high impact on trust as determined in the previous experiment. To measure the visual salience of each indicator form we asked participants after viewing the page to answer two verifiable questions that required recalling information from the notice and mention of the intervention. Two independent judges including the first author reviewed these answers to determine whether they showed evidence of the participant having seen the indicator, with disagreements resolved through discussion. Similar to experiment 1, trustworthiness assessments were solicited and their explanations were used to clarify any doubts regarding salience. All placements were performed on the ’Skyscraper’ article from the previous experiment as it was considered relatively uncontroversial and without distracting content.

4.1 Participants

Participants were recruited via Amazon’s Mechanical Turk service under the same conditions as the previous experiment. A total of 140 users participated, validated by Turkerview to exceed $15 an hour on average.

4.2 Results

In general, placement appeared to have a major influence on salience, with top performing indicator forms (Top-Right Notice (Blinking), Article-Content-Top Notice, Article-Content-Top Notice (Doubled-Height), and the Additional Citations article banner) at 55% saliency and others significantly lower, Warning Icon with Popup (36.8%), Top-Right Notice (33.3%), Notification Icon with Popup (30%). Motion also had a significant influence, Top-Right Notice (Blinking) outperformed Top-Right Notice by 67% with the addition of a slow blink. However, size appeared to have no effect: Article-Content-Top Notice performed very similarly to its Doubled-Height version, with functionally no difference (+0.6%) despite a footprint of double the size.

In summary, our results suggest that the most salient results can be achieved with right-hand placement and a bit of subtle motion. A number of approaches we explored that appeared to be reasonable places for indicators and that we expected to have greater effects were instead often not noticed. Another high salience factor was motion, which was found very noticeable but would likely need to be used sparingly or subtly in order to attract attention without becoming distracting. In Experiment 3, we build on these findings when designing an aggregate indicator, using the placement described above and including motion through a subtle one-shot animation that plays after page load, showing the indicator dial moving up through all levels then back down to the specified level.

These results also suggest that indicators may face a natural upper limit in regards to saliency. We hypothesize that this may be caused by user expectations of indicator placements - several participants explicitly noted their expectation that article issues be raised in a banner on top of article content. It is possible that users may cease scanning the page for indicators if they are unable to find them in expected locations. Further research would need to be conducted to determine the exact reasons behind this, and if this were to be the case, what locations are most commonly expected for such notices.

5 EXPERIMENT 3: DESIGNING AND TESTING AN AGGREGATE INDICATOR

While Experiment 2 addressed the challenge of designing a compact yet salient trust indicator, a remaining issue is that Experiment 1 only identified ways to negatively impact trust, with little evidence that even notices indicating high article quality metrics could positively impact trust. To address this we explore a design that aggregates several metrics into a glanceable indicator, testing whether such a design might ameliorate the potential negative perceptions driven by either unfamiliarity with individual notices or more general concerns raised about the system.

Specifically, we designed a new trust indicator that surfaces an aggregate trust metric and enables the reader to drill down to see component metrics which were contextualized to make them more understandable to an unfamiliar audience. We also aimed at taking the best practices from Experiment 2 to create a compact and glanceable indicator form that was made salient through selecting one of the highest performing placements and subtle use of one-shot motion on page load. We selected placement [B] because it was tied for most salient (with [A] Notice (Blinking)) and was most consistent with where notices are currently placed in Wikipedia, thus making it easier to support future deployments with less change to familiar page structures.

In order to create the indicator itself, we combined guidelines for trust visualization indicators from related literature as well as our prior experiment results, with an intention to cover factors

Placement Description	Position	N	Mean Salience
[A] Notice (Blinking)	Top-Right	18	0.556
[B] Notice	Top of Article Content	18	0.556
[C] Notice (Doubled Height)	Top of Article Content	20	0.55
[D] "Additional Citations" Template	Top of Article Content	20	0.55
[E] Warning Icon (Notice on click)	Immediately Left of Article Title	19	0.368
[F] Notice	Top-Right	18	0.333
[G] Notification Icon (Notice on click)	Right of 'Talk Page' tab	20	0.3

Table 4: Descriptions and effects of placements, number of participants, and salience ratings for notices tested in Experiment 2.

that could both highlight issues with low quality content yet also highlight positive aspects of high quality content. Specifically, in our indicator design (viewable in Figure 1), we apply these in the following ways:

- The use of an intuitive red-yellow-green color palette and the use of five levels to give "the best balance between abstraction, simplicity and detail" [67], resulting in a five-segment 'trustworthiness gauge'.
- A 'scoring explanation' on hover, influenced by literature that a set of in-direct perceptual cues designed from user feedback is proven to improve user acceptance of credibility decision aid recommendations [35] as well as our experiment results regarding the 'multiple issues' banner.
- Clear 'scoring factor explanations' beneath each 'scoring factor', supported by prior literature demonstrating that easy-to-understand decision process labels are important for improving trust perceptions of a recommendation or decision explanation [19].
- A leading 'Quality Rating' factor in the 'scoring factors', based on prior literature in credibility visualization suggesting a 'Competency' factor is a dominant influence in trustworthiness perceptions within visualizations that also display multiple other trust-related metrics [67].
- A brief 'start-up' animation of the gauge for increased salience in a low-footprint location, based on our findings from experiment 2.

The resulting gauge and scoring explanation system is designed to communicate a diverse set of trustworthiness calculations and values. The gauge has five color-coded segments, while the 'scoring explanation' panel can be mapped to a range of continuous or discrete values.

5.1 Experiment Design

To evaluate the new indicator we used a 2x2x6 design similar to Experiment 1 with two levels of quality and controversy and testing all five levels of indicator values and a no-indicator control. Since we aim to understand the relationships between article content, indicator value, and change in perceptions of trustworthiness, we aimed to find representative articles in Wikipedia which readers would have some understanding of the article content to potentially assess the accuracy of the indicator value shown. Thus, high and low quality articles were selected at random from the top 33% most popular Wikipedia pages by pageviews in 2019. Within this tertile, controversial articles were randomly selected from articles identified as controversial by the editor community using the 'Controversial' template. These were additionally validated for reader perception of subject controversy via a short pre-test survey with Amazon Mechanical Turk participants (n=15) using a 7-point scale. The articles selected are listed in Table 6. The average perceived controversy scores of each was as follows (higher is more controversial): -2.6 (Eagle), -1.47 (Metric System), 0.94 (Anarchism), and 1.2 (The Satanic Verses).

	High Quality	Low Quality
Controversial	Anarchism	The Satanic Verses
Uncontroversial	Metric System	Eagle

Table 5: Articles used in Experiment 3, sampled from differing levels of quality and perceived controversy.

In the experiment, levels of the gauge and scoring panel were manipulated to clearly communicate each segment of the gauge indicator, forming five conditions. In each condition, the panel elements are matched to the gauge by placing each 'scoring factor' in the respective scoring segment, with a small amount of visual jitter to increase the visual plausibility of results. Visual jittering was kept constant - values were randomly chosen once initially and then translated to the appropriate position. As a group, the scoring factors visually appeared in the middle of the respective segment. This 'scoring explanation' panel can be seen on the foreground of Figure 1. Additionally, there was a baseline control condition in which no indicator was shown.

5.2 Procedure

In addition to asking about local changes in perception for the article we applied the indicator to, building on our hypothesis from Experiment 1 we also asked participants about global changes in trust perceptions for Wikipedia as a whole. We collected user responses regarding changes in perceptions of article trustworthiness as well as Wikipedia overall in a two-stage approach similar to [85]. First, readers are shown the article and answer the content-related questions mentioned in Experiment 1. Afterwards, they are shown the same article, this time with the indicator inserted. Participants are directed near the article title and told that an indicator has been added "that shows the trustworthiness score of the article, calculated from publicly available information regarding the content of the article, edit activity, and editor discussions on the page". This is followed by a brief explanation of the scoring panel and its activation. After readers review the article a second time with the indicator, they are asked multiple questions relating to the content of the indicator as a manipulation check, such as the indicated gauge segment, the number of scoring factors, and the name of the fourth (from the top) scoring factor. Afterwards, they are asked to rate the effect of what they've been shown on their perceptions of Article and Wikipedia Overall Trust on a 7-point scale: ⟨Strongly, Moderately, Slightly⟩ x ⟨Raising, Lowering⟩ or 'Did not affect'.

Each participant was randomly assigned one article and an article indicator position. For the purposes of calculating a baseline, a no-intervention condition was assigned randomly as well. If participants were in this baseline condition, they were not shown a second page or asked any intervention-related questions. Article-related trust change questions were not asked in this condition.

5.3 Participants

Participants were recruited, compensated, and excluded as before. A total of 383 users participated.

5.4 Results

A key question we aimed to answer here was whether, in addition to reducing trust at low indicator levels, high indicator levels would respectively increase trust. Concerns about showing how "the sausage is made" suggest that an indicator that raises awareness of the contributor-driven nature of Wikipedia might result only in reductions in trust. However, we find reliable increases in trust at top indicator levels, averaging a +0.95 pt change. The effect of the top indicator levels was largely consistent among articles, with the exception of the Uncontroversial article regarding 'Eagles' (on average, Light Green produced a 0 pt change while Green produced a 1.89 pt change). It's possible this is due to high expectations regarding the trustworthiness of this common subject. This suggests that a trust indicator can provide system designers with the tools to dial trust in both positive and negative directions, under the assumption that designers choose accurate and representative mappings between indicator levels and article characteristics.

Decreasing levels of the indicator were associated with decreasing levels of perceived trust (see Figure 4, stats). However, decreases in trust were not uniform, with a relatively shallow decrease between dark green and light green, followed by a steep dropoff moving into yellow and below. These results are consistent with a threshold model in which the presence of even a small number of issues could lead to a wholesale discounting of the article's trustworthiness. This dropoff was especially pronounced for controversial articles (averaging a -1.58 pt change among the three 'untrustworthy' states), which might prime a reader's sensitivity to potential issues.

Diving further into participants' qualitative responses, we noticed that their reactions to the trust indicator seemed to be quite varied. Some said that they valued seeing details about the process by which the article was made, which gave them more confidence about making accurate trust judgments about the article:

Statistic	Coef.	SE	P-Value
Intercept	-1.492	0.196	<0.001
Indicator Level	0.494	0.064	<0.001
Seen a Discussion Page	1.038	0.377	0.001
Indicator Level X Seen a Discussion Page	-0.260	0.111	0.02

Table 6: Interaction Analysis Variables regarding changes in Wikipedia Trust in Experiment 3.

> "I like that it gives more information about how well the article is made and shows me whether or not i can trust the information given."
>
> "I think that having a metric to state the contentiousness of an article or the veracity of certain claims is a good thing. Transparency when it comes to information and where that information is coming from is very important."

However, others seemed to feel that being exposed to such process information violated their previous, largely positive, expectations of Wikipedia:

> "I felt that all content on this site was supposed to be trustworthy and now I have been shown otherwise."
>
> "This makes it look like they don't keep track of what is on there."

One factor that could explain this dichotomy of views is participants' previous experience with and expectations about Wikipedia. Those whose mental model of Wikipedia already incorporates the knowledge of its mutable and user-contributed nature might value interventions that surface information about that process and increase transparency. Conversely, those that had (incorrect) expectations, e.g., about Wikipedia as an organization that vetted all content before it was published, might feel less trusting once made aware of that faulty assumption.

To test this we conducted a regression analysis using an OLS model predicting change in trust from the independent variables of: indicator level, whether the participant reported having seen a Wikipedia discussion page, and their interaction. We found that both indicator value and seeing a discussion page were significant main effects, with those having been exposed to discussion page content having a more positive view of Wikipedia. Interestingly, there was a significant negative interaction between the two, suggesting that those more knowledgeable about Wikipedia's process may be less affected by lower indicator values. The interaction plot in Figure 6 provides confirmatory evidence of this, with those with discussion page experience less negative at indicator values 1 and 2 than those without such experience. Another interesting finding from the plot is that those without discussion page experience seem to be highly positively affected by both indicator values 4 and 5, while trust for those with experience drops off more sharply by indicator value 4. These results are suggestive that there may be different mental models at play for those with and without experience by which Wikipedia works, with those not knowing following an all-or-none model in which the presence of issues leads to a sharp threshold discounting of trust, whereas for those more familiar with Wikipedia's process trust is more linear and gradated. While

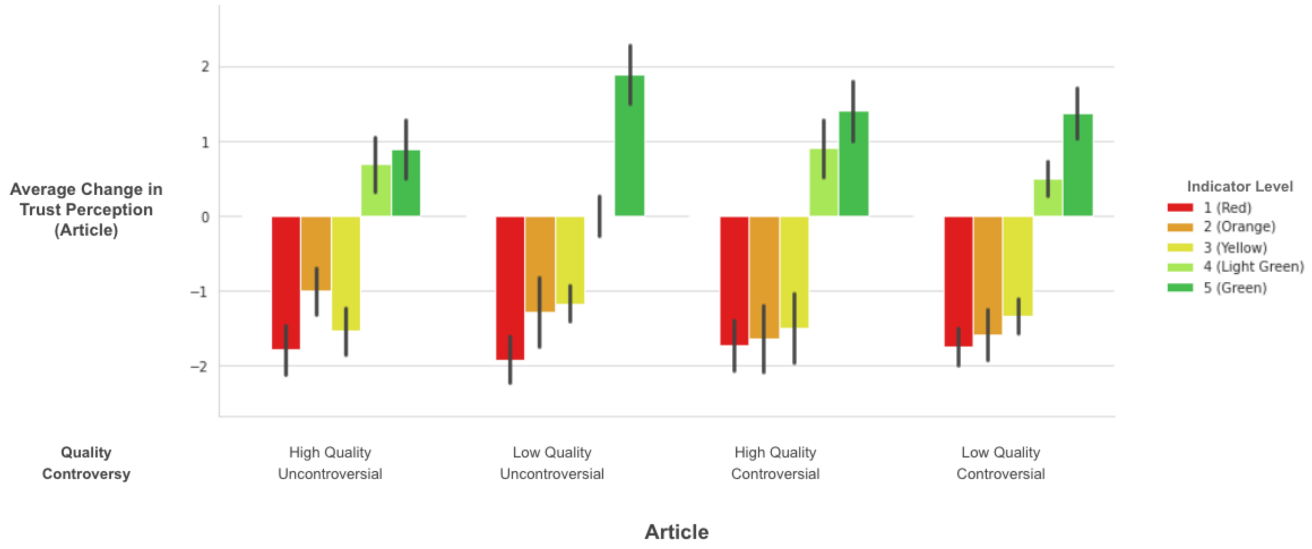


Figure 5: Impact of differing indicator values across article contexts on changes in article trust for the article the indicator was surfaced on. Lines correspond to standard error.

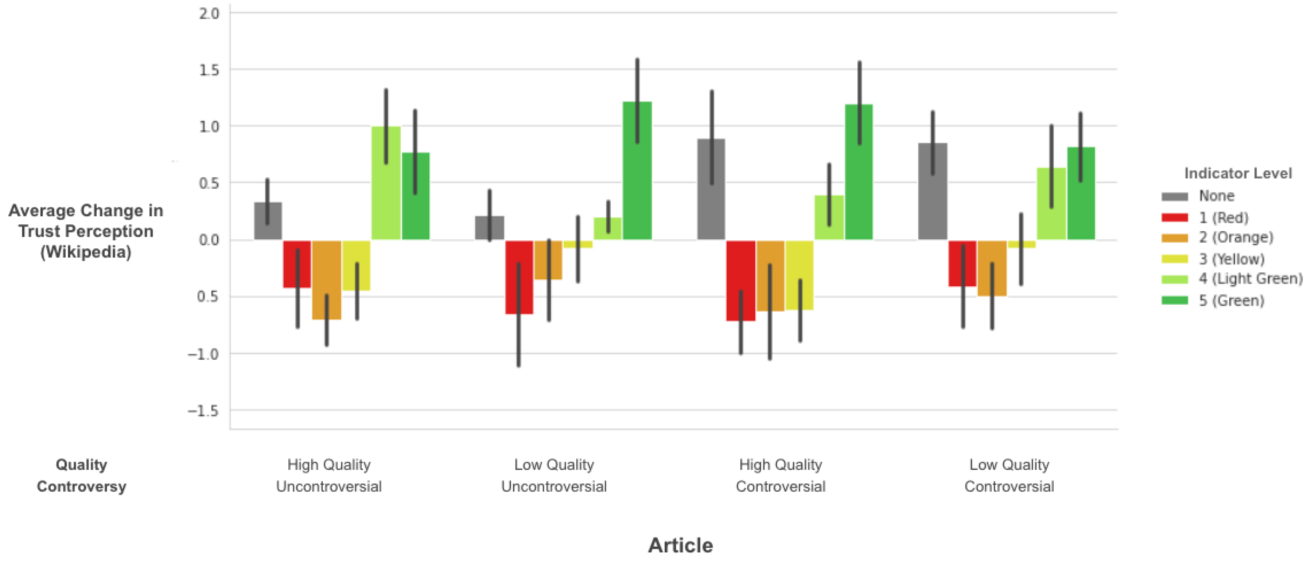


Figure 6: Impact of differing indicator values across article contexts on changes in overall Wikipedia trust. The 'None' indicator value (represented by gray bars) corresponds to the control condition with no indicator present (i.e., how seeing the article with no intervention added changed reader trust perceptions in Wikipedia overall). Lines correspond to standard error.

more research is needed to follow up on this intriguing finding, it suggests that designers may need to take into account how trust indicators may differentially affect readers with varying levels of knowledge of the production processes involved in user generated content communities.

6 DISCUSSION

In this paper we explored factors underlying the 'last mile' problem of creating trust indicators for user generated content in Wikipedia. Our work explored the design space for visual trust indicators, empirically investigating which signals would most affect perceptions of trust, and trade-offs between compactness and placement versus saliency of a potential indicator. In response to our initial findings,

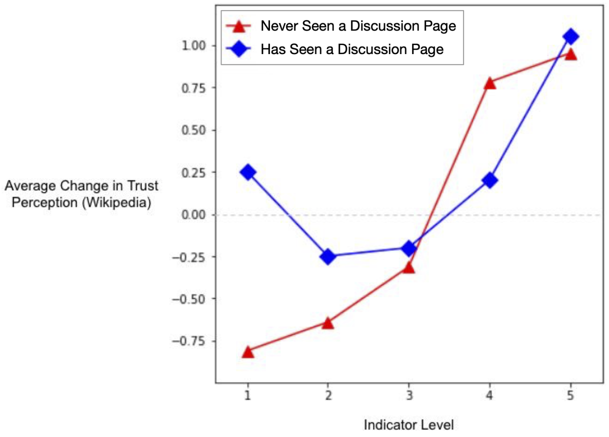


Figure 7: Interaction plot of indicator value (x-axis) versus average change in Wikipedia trust perceptions (y-axis). Participants unaware of Wikipedia discussion pages (red) show a steeper dropoff at low trust values compared to those familiar with them (blue), but less sensitivity at high trust values.

we created a sample trust indicator and demonstrated that surfacing authoring-related metrics to readers can positively, not just negatively, influence reader perceptions of trust. Furthermore, we characterized the response curve for each level of the indicator across multiple types of Wikipedia pages and showed that this effect significantly differs based on readers' prior knowledge of the content production process. Our results provide empirically-driven guidelines for the development of visual trust indicators that may help readers more accurately understand the quality and bias of user-generated content systems. Below we discuss additional considerations required for the development of such indicators on Wikipedia, as well as implications of our research for other user-generated content communities.

While this research probed some key questions for trust indicator design and deployment, there are several remaining degrees of freedom, parameters, decisions, and future research remaining for both system designers and researchers. A future system designer planning to implement a site-wide trust indicator (similar to that shown in Experiment 3) would need to make deliberate choices regarding the mapping between the indicator values and the pages they hope to affect. In particular, our results show very strong changes in trustworthiness as soon as any issues are brought to light. Implementing an effective and consistent system would therefore require researchers to work with the respective contributor community to assess which issues are significant to surface despite risking big drops in reader trust. One potential solution may involve community review of any indicator scoring, and perhaps even an implementation of 'community-in-the-loop'-style participatory design [81]. In the more immediate context, this thread of research could also help inform current Wikipedia editors as they negotiate the presence and placement of existing maintenance templates. As we have found in our work, some templates likely have high user trust impact and should be used strategically, while some, such as templates that use Wikipedia-terminology (e.g., 'Original Research') seem to have little impact on user trustworthiness perceptions and perhaps need to be improved. However, as previously mentioned, it would be very important for any system designers to triangulate with Wikipedia editors for their feedback on metrics, visual designs, and wordings used in such indicators as preferences will likely vary between more and less experienced editors, as well as vary by subject area or community (e.g. Wikiproject Medicine).

Our results suggested that some displayed metrics are much more impactful than others, however, in many cases those metrics would likely be difficult to reliably calculate with off-the-shelf natural language processing models. In order to do so, future implementations may have to introduce new sociotechnical systems and externalities to support such calculations. For example, input for variables such as 'Editor Disputed References' may come from a structured process of editor voting and 'Considered Complete By Editors' could be more easily computed if the format of Wikipedia article promotion discussions were augmented to include such text explicitly. Identifying impactful metrics could be positioned as part of identifying the machine learning needs of a given user-generated content community community and implemented as extensions to existing systems such as Wikipedia's ORES [28]. Using these automated methods as triggers to reduce the cost for editors to vet and correct incoming data could result in sustainable incentive structures for generating reliable metrics. Wikipedia's 'template warnings' are preliminary existence proofs of such structures, as well as bots that create warnings and flag material for editors to review. The high visibility of these signals on pages that are accessed by general readers could be a strong incentive for editors to keep metrics accurate and updated [93].

More generally, we see the sociotechnical architectures needed for reliable and sustainable indicator data as an opportunity for creating editor experiences that validate and communicate the value of contribution to both readers and editors. For example, a sample experience might highlight websites or search results where a specific piece of community content is embedded, which may nudge contributors to collaborate on clarifying the writing and re-examining the references. Furthermore, in the immediate context of Wikipedia, such indicators could increase reader engagement with Wikipedia's article content, such as references, that otherwise would have been hidden far down on the page [52]. Another possibility is using these signals to engage editors with the cumulative legacy of their individual contribution. For example, service awards (i.e., "barnstars" [45]) could be created to recognize editors for contributions to improving the accuracy of trust data in addition to existing awards for content and process contributions [92]. Additionally, these signals could support editors by enabling more nuanced recommender systems to help Wikipedia editors find useful pages to contribute to (e.g. [15]) or find editors for mentorship and socialization (e.g. [29], [63]).

The signals described in the paper could also help researchers better characterize the activities of editors and pages in other social production systems as well as the immediate context of Wikipedia. Signals such as debates among editors and open issues could be used to power new quantitative studies of the relationships between coordination, conflict, and quality (e.g., [41]) in such communities. In the context of Wikipedia research, our findings regarding familiarity with content production connects with prior work on the variety of roles editors take on and their activities within the larger Wikipedia community (e.g. [92], [11], [70]).

6.1 Beyond Wikipedia

Many other internet platforms might also benefit from compact, interpretable, and widely deployable signals of trust. As previously mentioned, user-generated content communities such as wikis generate many of the same signals as Wikipedia, suggesting that many of our results could be directly applicable, though the context and culture of the contributor community must be carefully considered. More generally, it can facilitate efforts to bridge the process of creating knowledge (which often leaves many process artifacts from annotators, editors, experts, etc) and its consumption, which is often divorced from such process information. Centralized peer production efforts, such as open source software and documentation projects, as well as more distributed and crowdsourced educational content, such as FOAM (Free Open Access Medical Education), could build trust by collaborating on community standards for trust-related indicators that would benefit readers' judgments of which content to consume and the confidence they should have in it. This could be implemented by embedding 3rd party HTML widgets, similar to "repository badges" currently used by Github projects to help users assess the state of a codebase. Such efforts could extend beyond initial creation to support more transparent curation as well. For example, news aggregators such as Google and Apple News could experiment with engaging users with the process by which editors or algorithms curate stories in order to increase user trust and address user misgivings regarding algorithmic curation. This could be similarly used in content aggregators such as Reddit to smoothen the experience of being exposed to poor quality content, such as providing additional context to 'Misleading Title" and similar flairs. In the context of social media, an implementation might include data from the chain of retweeters instead of Wikipedia editors, and placement in the Twitter feed instead of position on an article page (e.g. Twitter's infamous "fact-check labels" [76]). In the realm of journalism, this could involve surfacing information about individual articles, such as a small visual notice that an article is more than several years old or has been disputed in a more recent story [8]. Recently, a number of platforms have explored the use of reader-facing indicators as part of moderation efforts, and it is possible such interventions will become a ubiquitous face to content moderation where removals degrade the user experience [64]. We discuss this in the next section.

Lastly, our results regarding the potential for trust indicators to significantly lower trust in user-facing content contributes to existing models of content trust while simultaneously providing support for the use of content labeling as a potential complement to current methods of content moderation. The significant trust impacts we witnessed of notices that surfaced references, bias, and editor assessments of completeness track directly with the six factors found by researchers to be important in making content trust decisions of online resources: information provenance (related resources, provenance and pedigree), bias (perceived bias of the source, and perceived incentive to provide accurate information, likelihood of deceptive behavior), and recognized authority of the source writers [24]. Newer reviews of content trust models that aggregate literature from the fields of information quality and trustworthiness reveal many additional testable dimensions of provenance, perceived quality, and a priori user trust to be explored [68]. However, these models seem to treat trust factors as atomic in raising and lowering user trust, whereas our work demonstrates a potentially asymmetrical effect when these factors are mentioned positively or negatively. Focusing on this latter negative effect, the significant impacts of the negative notices tested in our studies also highlight the potential of trust indicators to contribute to improved content moderation strategies for other platforms. Wikipedia's approach of enabling community content creators to apply warnings to content is juxtaposed with the moderation processes of current social media websites, where users utilize individual 'reports' to advocate for content removal, which may then undergo judgement by a review process often opaque to users [66]. Prior literature on user consumption of labeled content highlights several benefits that Wikipedia users also likely experience. First, moderation-related explanations reduce frustration, provide the original poster valuable feedback, and have been shown to modify the future behaviors of users [39]. In the context of Wikipedia, labeling a past edit may offer a reframing of the revert process that helps retain junior editors that have committed a mistake or subpar edit. Additionally, labeling content instead of quietly removing it could avoid undesirable user perceptions of content removal, such as user confusion, frustration, and perceptions of unfairness [38]. Lastly, labeled posts create a passive resource for users that may help them learn community norms and thus perceive future moderation actions as fair [38], potentially even for controversial and political subjects [90]. In the context of Wikipedia, the trust impact results within our work empirically support the plausibility of implementing additional and more widespread content labeling within the platform: we demonstrate that there exist categories of issues that reliably and significantly lower content trust across diverse subject areas and demonstrate the potential for such indicators to have significant positive impacts on platform trust, even when indicating a critical view of the immediate article being viewed. Outside of Wikipedia, implementing content labeling for moderation may be done in a relatively straightforward fashion, for example creating features that enable users to design and attach reader-facing messages to content. As a more concrete example, a chat room (e.g. discord server) could enable users to apply engaging warning label 'message reactions' designed by the community underneath scam-related messages. Although our work did not explicitly explore this use case, we hope our work contributes to the ongoing conversation regarding the use of content labeling as an alternative to content removal in moderation processes.

6.2 Future Work

Finally, directly extending this work for Wikipedia content will require future research in several key areas including the impact of diverse content, sustainability of indicators, and the impact of diverse information consumers. Similarly, transferring the lessons

[8] At the time of this writing, Forbes Magazine implements such a notice consisting of a small red clock icon and the phrase 'This article is more than [X] years old' below the title.

from our work to other user-generated content communities will require additional exploration in the analogical topics for the intended platform.

Our experiments involved relatively familiar articles with substantial content; even for the "low quality" cases articles typically contained several images and more than 2,000 words. However many articles (by number, but not necessarily viewcount) are short, unfamiliar, and discuss niche subjects. This is an exciting area of potential impact as prior literature implies that readers of unfamiliar subjects desire the assistance of a trust support system and accept its recommendations more frequently [55]. In addition, there is also a natural variety in the reasons articles are controversial that may affect their perception which we did not investigate here. For scope reasons we chose articles that were representative of the most commonly encountered Wikipedia articles. However, we did not investigate very long and thorough articles, many of which are some of the most popular subjects on Wikipedia, in part due to participant time constraints. In such cases it would be interesting to study the effect of section-level banners and their various permutations. It is also possible that our indicator could have a stronger effect on such articles, as it simplifies and aggregates notices that might be visually distant from one another and easy to miss. Our experience in the sampling process suggested to us that articles rated as relatively low quality (e.g., C-class) by Wikipedians still had a substantial number of edits and word count and might reasonably generalize from our sampled articles. It is less clear how Stub and Start class articles would fare as they have fewer historical signals to surface, which we believe is an important topic for future research. In the larger context of user-generated content communities, this could take the form of exploring the impact of content at both extremes of acceptable length, for example a tweet-sized document or long post written by a passionate power-user. One option for articles or content with partial or unreliable signals is to explore an "in progress" indicator which would indicate uncertainty rather than negative valence. Similar to our scoping constraints with pages, for scope reasons we were unable to test all existing Wikipedia banners and instead tested the most popular issue-related banners. Thus, understanding the impact of rarer and non-issue-related banner templates such as navigation is left as important future work. In particular, the rarer but often viewed banners regarding 'recent or ongoing' events are good targets for future study, as pages often experience a sudden burst of editing activity due to a recent event [23] and the reader trust impact of editors placing the 'recent events' banners are likely dependent on topic domain and external context. An interesting research direction here could investigate the types of real-world events (or subsequent pro-social activities [82]) that stimulate such editing activity and develop frameworks that generalize across the many different topic areas of Wikipedia.

Indicator designers will also face the prospect of designing for an audience with diverse individual experiences (or perhaps even affiliation with a user-generated sub-community) in their platform. Our work used US-based AMT workers as research participants to represent Wikipedia's audience, which may have limitations when generalizing to a global viewpoint. However, studies of AMT demographics suggest that they are in many ways similar to the general population [34][49][72], with some caveats (e.g., younger, more technologically savvy) that may also reflect trends in Wikipedia's demographics [32] [26]. All of our AMT workers were familiar with and had read Wikipedia pages. Despite this, our data do not rule out the possibility that segments of the population not included in our sampling distribution may behave differently.

In this work we examined several key factors regarding the 'last mile' problem of creating trust indicators for user generated content, situating our focus on one of the largest of such sites, Wikipedia. Our results explored the design space of such indicators - and investigated the effect of signals on user perceptions of trust, trade-offs between compactness and placement versus saliency of an indicator, and characterized the response curve for reader trust perceptions of a sample trust indicator. More generally, if successful such indicators could kickstart processes to combat misinformation across a variety of online sources by engaging readers and motivating editors in the difficult but vital work of collaborative knowledge production and curation.

ACKNOWLEDGMENTS

This work was supported by NSF grants FW-HTF-RL-1928631, CCF-1814826, IIS-1816242, the Office of Naval Research, the Carnegie Mellon University Center for Knowledge Acceleration, and the Wikimedia Foundation. We'd also like to thank the reviewers for their generous review and feedback. Finally, we'd also like to acknowledge our study participants, without whom this work would not have been possible.

REFERENCES

[1] 2022. Wikipedia Pageview statistics. https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics

[2] B Adler, Luca De Alfaro, and Ian Pye. 2010. Detecting wikipedia vandalism using wikitrust. *Notebook papers of CLEF* 1 (2010), 22–23.

[3] B Thomas Adler. 2012. *WikiTrust: content-driven reputation for the Wikipedia.* Ph.D. Dissertation. UC Santa Cruz.

[4] B Thomas Adler, Krishnendu Chatterjee, Luca De Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning trust to Wikipedia content. In *Proceedings of the 4th International Symposium on Wikis.* 1–12.

[5] B Thomas Adler and Luca De Alfaro. 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web.* 261–270.

[6] Yochai Benkler. 2017. Peer production, the commons, and the future of the firm. *Strategic Organization* 15, 2 (2017), 264–274.

[7] Jan Panero Benway. 1998. Banner blindness: The irony of attention grabbing on the World Wide Web. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 42. SAGE Publications Sage CA: Los Angeles, CA, 463–467.

[8] Erik Borra, David Laniado, Esther Weltevrede, Michele Mauri, Giovanni Magni, Tommaso Venturini, Paolo Ciuccarelli, Richar Rogers, and Andreas Kaltenbrunner. 2015. A platform for visually exploring the development of Wikipedia articles. In *Ninth International AAAI Conference on Web and Social Media.* Citeseer.

[9] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal controversies in Wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* 193–196.

[10] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, Tommaso Venturini, et al. 2014. Contropedia-the analysis and visualization of controversies in Wikipedia articles.. In *OpenSym.* 34–1.

[11] Susan L Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work.* 1–10.

[12] Erik Brynjolfsson, Felix Eggers, and Avinash Gannamaneni. 2018. Measuring welfare with massive online choice experiments: A brief introduction. In *AEA Papers and Proceedings*, Vol. 108. 473–76.

[13] Fanny Chevalier, Stéphane Huot, and Jean-Daniel Fekete. 2010. Wikipediaviz: Conveying article quality for casual wikipedia readers. In *2010 IEEE Pacific Visualization Symposium (PacificVis).* IEEE, 49–56.

[14] Si-Chi Chin, W Nick Street, Padmini Srinivasan, and David Eichmann. 2010. Detecting Wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on Information credibility*. 3–10.

[15] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 32–41.

[16] Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. 27–30.

[17] Peter Denning, Jim Horning, David Parnas, and Lauren Weinstein. 2005. Wikipedia risks. *Commun. ACM* 48, 12 (2005), 152–152.

[18] Pierpaolo Dondio, Stephen Barrett, Stefan Weber, and Jean Marc Seigneur. 2006. Extracting trust from domain analysis: A case study on the wikipedia project. In *International Conference on Autonomic and Trusted Computing*. Springer, 362–373.

[19] Kimberly D Elsbach and Greg Elofson. 2000. How the packaging of decision explanations affects perceptions of trustworthiness. *Academy of Management Journal* 43, 1 (2000), 80–89.

[20] Andrew J Flanagin and Miriam J Metzger. 2011. From Encyclopaedia Britannica to Wikipedia: Generational differences in the perceived credibility of online encyclopedia information. *Information, Communication & Society* 14, 3 (2011), 355–374.

[21] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.

[22] Brian J Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 80–87.

[23] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. 2013. Extracting event-related information from article updates in wikipedia. In *European Conference on Information Retrieval*. Springer, 254–266.

[24] Yolanda Gil and Donovan Artz. 2007. Towards content trust of web resources. *Journal of Web Semantics* 5, 4 (2007), 227–239.

[25] Jim Giles. 2005. Internet encyclopaedias go head to head.

[26] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group* 8 (2010), 1158–1178.

[27] Shane Greenstein and Feng Zhu. 2018. Do experts or crowd-based models produce more bias? Evidence from encyclopædia britannica and wikipedia. *Mis Quarterly* (2018).

[28] Aaron Halfaker and R Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *arXiv preprint arXiv:1909.05189* (2019).

[29] Aaron Halfaker, Bryan Song, D Alex Stuart, Aniket Kittur, and John Riedl. 2011. NICE: Social translucence through UI intervention. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. 101–104.

[30] Stephen Harrison. 2020. Twitter wants to use wikipedia to help determine who gets a blue checkmark. https://slate.com/technology/2020/12/twitter-checkmark-verification-wikipedia-notability.html

[31] Marit Hinnosaar, Toomas Hinnosaar, Michael E Kummer, and Olga Slivko. 2019. Wikipedia matters. *Available at SSRN 3046400* (2019).

[32] Paul Hitlin. 2016. Turkers in this canvassing: young, well-educated and frequent users. *Pew Research Center* 437 (2016).

[33] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 243–252.

[34] Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010).

[35] Matthew L Jensen, Paul Benjamin Lowry, and Jeffrey L Jenkins. 2011. Effects of automated and participative decision support in computer-aided credibility assessment. *Journal of Management Information Systems* 28, 1 (2011), 201–234.

[36] Grace YoungJoo Jeon and Soo Young Rieh. 2014. Answers from the crowd: how credible are strangers in social Q&A? *IConference 2014 Proceedings* (2014).

[37] Johan Jessen and Anker Helms Jørgensen. 2012. Aggregated trustworthiness: Redefining online credibility through social validation. *First Monday* (2012).

[38] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[39] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[40] Kari Kelton, Kenneth R Fleischmann, and William A Wallace. 2008. Trust in digital information. *Journal of the American Society for Information Science and Technology* 59, 3 (2008), 363–374.

[41] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 37–46.

[42] Aniket Kittur, Bongwon Suh, and Ed H Chi. 2008. Can you ever trust a Wiki? Impacting perceived trustworthiness in Wikipedia. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 477–480.

[43] Vasilis Kostakis and Marios Papachristou. 2014. Commons-based peer production and digital fabrication: The case of a RepRap-based, Lego-built 3D printing-milling machine. *Telematics and Informatics* 31, 3 (2014), 434–443.

[44] Mark Kramer, Andy Gregorowicz, and Bala Iyer. 2008. Wiki trust metrics based on phrasal analysis. In *Proceedings of the 4th International Symposium on Wikis*. 1–10.

[45] Travis Kriplean, Ivan Beschastnikh, and David W McDonald. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 47–56.

[46] Sanne Kruikemeier and Sophie Lecheler. 2018. News consumer perceptions of new journalistic sourcing techniques. *Journalism Studies* 19, 5 (2018), 632–649.

[47] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*. 857–866.

[48] Zui Chih Lee and Jennifer Yurchisin. 2011. The impact of website attractiveness, consumer-website identification, and website trustworthiness on purchase intention. *International Journal of Electronic Customer Relationship Management* 5, 3-4 (2011), 272–287.

[49] Kevin E Levay, Jeremy Freese, and James N Druckman. 2016. The demographic and political composition of Mechanical Turk samples. *Sage Open* 6, 1 (2016), 2158244016636433.

[50] Sook Lim. 2013. College students' credibility judgments and heuristics concerning Wikipedia. *Information Processing & Management* 49, 2 (2013), 405–419.

[51] Teun Lucassen, Rienco Muilwijk, Matthijs L Noordzij, and Jan Maarten Schraagen. 2013. Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 254–264.

[52] Teun Lucassen, Matthijs L Noordzij, and Jan Maarten Schraagen. 2011. Reference blindness: The influence of references on trust in Wikipedia. (2011).

[53] Teun Lucassen and Jan Maarten Schraagen. 2010. Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*. 19–26.

[54] Teun Lucassen and Jan Maarten Schraagen. 2011. Evaluating WikiTrust: A trust support tool for Wikipedia. *First Monday* (2011).

[55] Teun Lucassen and Jan Maarten Schraagen. 2012. The role of topic familiarity in online credibility evaluation support. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. SAGE Publications Sage CA: Los Angeles, CA, 1233–1237.

[56] Brendan Luyt, Tay Chee Hsien Aaron, Lim Hai Thian, and Cheng Kian Hong. 2008. Improving Wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology* 59, 2 (2008), 318–330.

[57] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Eleventh International AAAI Conference on Web and Social Media*.

[58] Ericka Menchen-Trevino and Eszter Hargittai. 2011. YOUNG ADULTS'CREDIBILITY ASSESSMENT OF WIKIPEDIA. *Information, Communication & Society* 14, 1 (2011), 24–51.

[59] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.

[60] Jason Mittell. 2012. Wikis and participatory fandom. In *The participatory cultures handbook*. Routledge, 53–60.

[61] Jarkko Moilanen and Tere Vadén. 2013. 3D printing community and emerging practices of peer production. *First Monday* (2013).

[62] Jonathan Morgan and Dario Taraborelli. 2019. Research directions towards the Wikimedia 2030 strategy. https://wikimediafoundation.org/news/2019/02/14/research-directions-towards-the-wikimedia-2030-strategy/

[63] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 839–848.

[64] Garrett Morrow, Briony Swire-Thompson, Jessica Polny, Matthew Kopec, and John Wihbey. 2020. The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation. *Available at SSRN* (2020).

[65] Sai T Moturu and Huan Liu. 2009. Evaluating the trustworthiness of Wikipedia articles through quality and credibility. In *Proceedings of the 5th international symposium on wikis and open collaboration*. 1–2.

[66] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[67] Jason RC Nurse, Sadie Creese, Michael Goldsmith, and Koen Lamberts. 2012. Using information trustworthiness advice in decision making. In *2012 Workshop on Socio-Technical Aspects in Security and Trust*. IEEE, 35–42.

[68] Jason RC Nurse, Syed Sadiqur Rahman, Sadie Creese, Michael Goldsmith, and Koen Lamberts. 2011. Information quality and trustworthiness: A topical state-of-the-art review. (2011).

[69] Samantha R Paige, Janice L Krieger, and Michael L Stellefson. 2017. The influence of eHealth literacy on perceived trust in online health communication channels and sources. *Journal of health communication* 22, 1 (2017), 53–65.

[70] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*. 51–60.

[71] Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. 2020. Towards detection of subjective bias using contextualized word embeddings. In *Companion Proceedings of the Web Conference 2020*. 75–76.

[72] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188.

[73] Jarutas Pattanaphanchai, Kieron O'Hara, and Wendy Hall. 2013. Trustworthiness criteria for supporting users to assess the credibility of web information. In *Proceedings of the 22nd International Conference on World Wide Web*. 1123–1130.

[74] Peter Pirolli, Evelin Wollny, and Bongwon Suh. 2009. So you know you're getting the best possible information: a tool that increases Wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1505–1508.

[75] Russell A Poldrack, Aniket Kittur, Donald Kalar, Eric Miller, Christian Seppa, Yolanda Gil, D Stott Parker, Fred W Sabb, and Robert M Bilder. 2011. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in neuroinformatics* 5 (2011), 17.

[76] Nicholas Reimann. 2020. Twitter fact-checks President Trump for the first time. https://www.forbes.com/sites/nicholasreimann/2020/05/26/twitter-fact-checks-president-trump-for-the-first-time/?sh=4a98a4612905

[77] Antonio J Reinoso, Rocıo Munoz-Mansilla, Israel Herraiz, and Felipe Ortega. 2012. Characterization of the Wikipedia traffic. In *ICIW 2012: Seventh International Conference on Internet and Web Applications and Services*. 156–162.

[78] Elena Rocco. 1998. Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 496–502.

[79] Camille Roth, Dario Taraborelli, and Nigel Gilbert. 2008. Measuring wiki viability: an empirical assessment of the social dynamics of a large sample of wikis. In *Proceedings of the 4th International Symposium on Wikis*. 1–5.

[80] Diego Saez-Trumper. 2019. Online disinformation and the role of wikipedia. *arXiv preprint arXiv:1910.12596* (2019).

[81] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[82] Rebecca Solnit. 2010. *A paradise built in hell: The extraordinary communities that arise in disaster*. Penguin.

[83] Bongwon Suh, Ed H Chi, Aniket Kittur, and Bryan A Pendleton. 2008. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1037–1040.

[84] Marsha Ann Tate. 2018. *Web wisdom: How to evaluate and create information quality on the Web*. CRC Press.

[85] 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*. 1059–1068.

[86] Victor A Tsygankov. 2004. Evaluation of website trustworthiness from customer perspective, a framework. In *Proceedings of the 6th international conference on Electronic commerce*. 265–271.

[87] Dirk Van Der Linden, Emma Williams, Joseph Hallett, and Awais Rashid. 2020. The impact of surface features on choice of (in) secure answers by Stackoverflow readers. *IEEE Transactions on Software Engineering* (2020), 1–1.

[88] Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[89] Simon Waldman. 2004. Who knows. *The Guardian* 26 (2004).

[90] John Wihbey, Garrett Morrow, Myojung Chung, and Mike Peacey. 2021. The Bipartisan Case for Labeling as a Content Moderation Method: Findings from a National Survey. *Available at SSRN* (2021).

[91] Thomas Wöhner and Ralf Peters. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. 1–10.

[92] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10.

[93] Heng-Li Yang and Cheng-Yu Lai. 2010. Motivations of Wikipedia content contributors. *Computers in human behavior* 26, 6 (2010), 1377–1383.

[94] Tae Yano and Moonyoung Kang. 2016. *Taking advantage of wikipedia in natural language processing*. Technical Report. Technical report, Carnegie Mellon University Language Technologies Institute.

[95] Honglei Zeng, Maher A Alhossaini, Li Ding, Richard Fikes, and Deborah L McGuinness. 2006. *Computing trust from revision history*. Technical Report. Stanford Univ Ca Knowledge Systems LAB.