Protection from Evil and Good: The Differential Effects of Page Protection on Wikipedia Article Quality

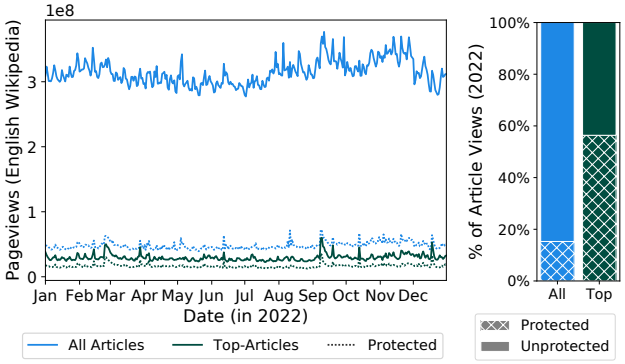
Thorsten Ruprechter,[1] Manoel Horta Ribeiro,[2] Robert West,[3] Denis Helic[1]
[1]Graz University of Technology
[2]Princeton University
[3]EPFL
th.ruprechter@gmail.com, manoel@cs.princeton.edu, robert.west@epfl.ch, dhelic@tugraz.at

Abstract

Wikipedia, the Web's largest encyclopedia, frequently faces content disputes or malicious users seeking to subvert its integrity. Administrators can mitigate such disruptions by enforcing "page protection" that selectively limits contributions to specific articles to help prevent the degradation of content. However, this practice contradicts one of Wikipedia's fundamental principles—that it is open to all contributors—and may hinder further improvement of the encyclopedia. In this paper, we examine the effect of page protection on article quality to better understand whether and when page protections are warranted. Using decade-long data on page protections from the English Wikipedia, we conduct a quasi-experimental study analyzing pages that received "requests for page protection"—written appeals submitted by Wikipedia editors to administrators to impose page protections. We match pages that indeed received page protection with similar pages that did not and quantify the causal effect of the interventions on a well-established measure of article quality. Our findings indicate that the effect of page protection on article quality depends on the characteristics of the page prior to the intervention: high-quality articles are affected positively, as opposed to low-quality articles that are impacted negatively. Subsequent analysis suggests that high-quality articles degrade when left unprotected, whereas low-quality articles improve. Overall, with our study, we outline page protections on Wikipedia and inform best practices on whether and when to protect an article.

1 Introduction

Wikipedia, one of the largest repositories of knowledge on the Web, serves the diverse information needs of users around the world (Lemmerich et al. 2019). From general topics, such as science and history (Singer et al. 2017), to events, such as pandemics (Ruprechter et al. 2021) and natural disasters (Lorini et al. 2020), the world turns to the online encyclopedia for fast, reliable, and free information. Wikipedia is made possible due to the work of volunteers (*Wikipedians*) who create new articles and curate existing content. At the core of Wikipedia's philosophy is its openness: "Anyone—including you—can become a Wikipedian by boldly making changes when they find something that can be added or improved (Wikipedia 2023b)."



(a) Time series of pageviews in 2022. (b) Aggregated.

Figure 1: Page protected articles correspond to a substantial fraction of page views Wikipedia receives. In (a), we show the daily number of views in 2022 that went towards: all English Wikipedia articles (blue), the 1,000 top-viewed articles per day (dark green), and page protected articles in both of these sets (dotted lines). From (b), we derive that, over the whole year, approximately 15% of overall views and 56% of views to the 1,000 top articles went to protected articles.

While most contributions to Wikipedia are constructive, good-faith editors may find themselves embroiled in edit wars about verbiage or conflicts about the validity of information (Faulkner, Walling, and Pinchuk 2012; Sumi, Yasseri et al. 2011); and malicious users might vandalize pages or manipulate the tone of an article through sockpuppetry (Kittur et al. 2007; Kumar et al. 2017), undermining the development of content and the integrity of knowledge on Wikipedia (Aragón and Sáez-Trumper 2021).

Regardless of users' intent, when disruptive editing patterns emerge, Wikipedia admins frequently intervene using "page protection" to safeguard them against further damage (Burke and Kraut 2008; Hill and Shaw 2015). This mechanism restricts who can edit an article temporarily (e.g., one week) or indefinitely and can be triggered by admins, either at their own will or when fulfilling user requests to protect a page (Wikipedia 2022a). Although less than 1% of pages in the English Wikipedia are protected at any given time (Hill and Shaw 2015; Spezzano, Suyehira, and Gundala

2019), protected articles receive a disproportionate amount of views (e.g., 15% of all views in 2022; see Fig. 1).

While page protections can effectively curb disruptive behavior, they are at odds with Wikipedia's openness principle, and thus, internal policies advocate using them scarcely (Wikipedia 2023a). They can deter newcomers (Halfaker et al. 2013) or alter the composition of editors working on particular content (Shi et al. 2019). Further, the rationale for protection is not always apparent (Das, Lavoie, and Magdon-Ismail 2016), leading some seasoned editors to regard them as harmful (Wikipedia 2021).

Present work. Given that page protections may be either harmful (preventing productive contributions) or beneficial (preventing disruptive editing), we analyze their usage and impact on the English Wikipedia. Specifically, we ask:

RQ1: How are page protections enacted on Wikipedia?

RQ2: What is the effect of protections on article quality?

For this, we first obtain 299 721 edit protections on the English Wikipedia between October 2012 and March 2023 using a methodology introduced by Hill and Shaw (2015). Second, we collect 127,098 user requests for page protection submitted during this period, along with their associated articles. Third, we enhance article revisions with metadata, such as article quality (Halfaker and Geiger 2020), which enables us to create a time series of metrics for every article.

Using these diverse data sources, we first characterize the usage of page protections on the English Wikipedia (**RQ1**). We then conduct a quasi-experimental study (Hernán and Robins 2016) to estimate the causal effect of page protections on article quality (**RQ2**). We match articles that were protected after a user's request for page protection to similar articles with a declined request. Using a dynamic difference-in-differences design, we then estimate the effect of protection on article quality, based on a well-established automated quality metric that considers features about article structure and text (Halfaker and Geiger 2020).

Findings. First, analyzing page-protected articles (**RQ1**), we observe that protections are typically brief, often lasting for a week or less, with certain pages being repeatedly protected over the course of several years. Edits generally increase before the enforcement of protections and decrease after the protection period concludes. Second, through our difference-in-differences analysis, we discover that page protections' impact strongly depends on the article's nature in the pre-intervention period (**RQ2**). Page protections are beneficial for high-quality articles, but detrimental to low-quality ones. Subsequent analysis indicates that this could be because protections safeguard high-quality content (potentially by curbing vandalism) but prevent low-quality articles from improving (potentially by curbing contributions). Furthermore, the effect varies across topics, with a more pronounced effect in articles related to "Geography" and "History & Society," a weaker effect in articles about "Culture," and no notable effect on "STEM" articles.

Implications. Our research offers valuable insights for editors and administrators of Wikimedia projects and other interested researchers. Specifically, our findings indicate that

Figure 2: **Example of a request that led to page protection.** In June 2018, the article "Thor (Marvel Comics)" was subject to disruptive editing because of increased attention, which regular editors (grey) had to clean up through reverts. Consequently, a user (blue) requested indefinite semi-protection to address the persistent vandalism by anonymous editors. Upon review, an admin (dark green) ruled against indefinite protection but still imposed a one-month semi-protection, highlighted on the article page through the "silverlock." An admin could have also declined the request (i.e., not protecting the page) or taken alternative actions (e.g., blocking users). (Screenshots taken from Wikipedia's requests for page protection archive at https://w.wiki/7EKd.)

the quality and topic of an article mediate the effect of page protection on article quality, which could be explicitly considered when deciding whether to protect an article or not. These findings can assist Wikipedia's stakeholders in making informed decisions and developing guidelines for page protection, promoting the continued enhancement of the largest encyclopedia on the Web.

2 Background: Page Protections

We now provide some background on page protection and requests for page protection on the English Wikipedia.

Protection process. Wikipedia is intended to be kept as open as possible (Wikipedia 2023a). Thus, admins are generally urged to enforce protection only when absolutely necessary to prevent further damage. Preemptive protection is generally not permitted (e.g., expected traffic due to current events), although certain articles are treated with greater caution (Wikipedia 2023a). However, in cases of disruptive editing behavior by multiple editors, Wikipedia administrators may be compelled to implement some form of edit pro-

	Access to Protected Wikipedia Pages by User Type or Role					
	Unregistered/New	**(Auto-)Confirmed**	**Extended confirmed**	**Template editor**	**Admin**	**Interface admin**
No protection	✓	✓	✓	✓	✓	✓
Pending changes	✪	✓	✓	✓	✓	✓
Semi	✗	✓	✓	✓	✓	✓
Extended confirmed	✗	✗	✓	✓	✓	✓
Template	✗	✗	✗	✓	✓	✓
Full	✗	✗	✗	✗	✓	✓
Interface	✗	✗	✗	✗	✗	✓

✗ = Can not edit; ✓ = Can edit; ✪ = Can edit, but changes are hidden from readers who are not logged in, until approved by a reviewer (logged in readers see changes right away). Table adapted from Wikipedia:Protection_policy (https://w.wiki/7QUz).

Table 1: **Page protection levels on the English Wikipedia.** For the most common levels of edit protection, the editor's account age and number of edits determine whether they can make revisions to an article. In the case of *semi-protection* (*extended confirmed protection*), editors must be *confirmed* (*extended confirmed*) by having an account for at least 4 (30) days and making at least 10 (500) edits. Pending changes protection is a special kind of extension of the Wikimedia software. This protection mode allows anyone to edit an article, but changes made by unregistered or unconfirmed editors must be approved by a pending changes reviewer before they become visible to readers who are not logged in. Higher protection levels (i.e., template, full, or interface) can only be bypassed by high-privilege user roles (template editor, admin, or interface admin).

tection.[1] The most commonly enforced protection is *semi-protection*, which requires an editor to have an account for at least four days and to have made at least ten edits. Stricter levels of protection exist, such as *full protection*, which only allows admins to edit content.

The level of protection enforced for an article depends on several factors, including the severity of the disruptive behavior, previous controversies, and past protections. Table 1 illustrates the six protection levels in the English Wikipedia as of September 2023. Admins enforce protection for either a temporary period, after which the article will be automatically unprotected, or for an indefinite period, which requires manual removal of the protection. Additionally, admins can adjust the protection levels during active protection.

Requests for page protection. While admins can enforce page protections independently, users can also submit requests for page protection (RfPP). When an RfPP is submitted through the request form, it automatically creates an entry on the RfPP overview page, which admins then review (Wikipedia 2022a). After review, admins make judgments through responses elicited by coded templates, such as *Declined*, *Semi-protected* or *Fully protected* (i.e., the request has been approved), *User(s) blocked*, *Already protected*, and others (for a full list, see https://w.wiki/6oRr). A single RfPP can have multiple responses from the same or different admins. For example, admins often seek additional comments and clarification from the requesting user or change their response within minutes of the initial decision. Once the admin's decision resolves the request, Wikipedia archives the RfPP on pages that are structurally similar to article talk pages (Wikipedia 2022b; Viégas et al. 2007).

Fig. 2 illustrates this procedure for an accepted RfPP. In this case, the article "Thor (Marvel Comics)" experienced vandalism from multiple anonymous editors, resulting in a user submitting an RfPP. Although the admin rejected the request for indefinite semi-protection, they decided to impose a temporary one-month semi-protection instead. Afterward, the "silverlock" signaling semi-protection appeared on the article page for "Thor (Marvel Comics)".

Protection in other languages. We focus on the English Wikipedia. While page protections and RfPP exist in other language editions, differences occur (Johnson and Lescak 2022). For example, in German, only four protection types exist (instead of six in English), and users do not request protection but instead report user misconduct (see https://w.wiki/7Rqj), making it difficult to compare language editions.

3 Related Work

Wikipedia content policing. Wikipedia uses a variety of policies and moderation tools to ensure productive user collaboration. First, admins can permanently block users for severe or repeated offenses (Das, Lavoie, and Magdon-Ismail 2016). Second, edit filters are mechanisms that detect whether contributions to a Wikipedia page violate certain criteria, such as whether they contain swear words or only use uppercase letters (Vaseva and Müller-Birn 2020). Depending on their configuration, the user will receive a warning, the edit summary of the revision will be flagged in the article revision, or the revision will be denied altogether. Third, many language versions use bots to automatically police content (Zheng et al. 2019). For example, *ClueBotNG*—one of the most notable bots in the English Wikipedia—automatically reverts vandalism (Wikipedia 2010). Certain tools (such as bots), as well as researchers, use the machine learning framework *ORES* to evaluate Wikipedia content and editor activity (Halfaker and Geiger 2020). ORES provides endpoints for several machine learning models that can be used to predict both article characteristics (e.g., topic or quality) and edit characteristics (e.g., whether it is damaging or made in good faith). In general, while admins attempt to sanction rogue users and vandalism swiftly, controversies can arise from valid changes, and their reversal can lead to edit wars (Yasseri and Kertész 2013).

Effects of page protections on Wikipedia. Early Wikipedia studies theorized rising page protections as a possible indication of increasing administrative action (Suh et al. 2009).

[1] While other protections exist (e.g., *move* or *create* protection), we focus on edit protection.

Furthermore, page protection can present additional obstacles for newcomers who already have difficulty integrating into the editorial community (Halfaker et al. 2013). Hill and Shaw (2015) describe the relevance of page protections and propose an approach to collect accurate accounts of protection "spells" for the English Wikipedia. They find that while only about 0.67% of articles on the English Wikipedia were protected in December 2013, these articles accounted for about 14.3% of all views on the online encyclopedia. Most relevant to our work, a recent study (Ajmani, Vincent, and Chancellor 2023) used a combination of qualitative and quantitative methods to assess the impact of protections in a single article category ("Internet Culture"). For these specific articles, protections lead to high editor turnover and requests can be categorized by editor activity, article topic, and article visibility. Overall, the literature to date has proposed data collection approaches and conducted parsimonious analyses of the impact of page protections on editors.

Predicting page protections. The process of manually enforcing page protection is tedious for both requesting users and admins. Thus, Spezzano, Suyehira, and Gundala (2019) built a machine learning classifier to predict whether a page will (or should) be protected. Their classifiers suggest that the most important features for classifying articles as worthy of protection are the rapidity of contributions (mean time between revisions), anonymous edits, and article topic.

Protection as a proxy of managerial authority. DeDeo (2016) leverages page protections as an example of top-down authority, which is driven by reverts. His analysis (62 articles) suggests that both the start and end of protections rarely have lasting effects on long-term activity. Similarly, Klapper and Reitzig (2018) view page protection as enforcement of lateral authority and find that affected editors decrease their activity on talk pages of protected articles but increase their activity elsewhere.

Restrictions in other Web platforms. Other community-driven websites employ mechanisms comparable to page protections. On the discussion platform Reddit, admins possess the ability to restrict (i.e., only selected users can create new posts), quarantine (i.e., hide on Reddit's search and front page), or completely ban sub-forums. This has been associated with a decline in activity and migration of users to alternative platforms (Horta Ribeiro et al. 2021; Chandrasekharan et al. 2017). In contrast, automated content moderation on Facebook has been shown to promote adherence to rules without decreasing commenting activity (Horta Ribeiro, Cheng, and West 2022).

Contributions of our work. We extend these previous works by (i) providing code to retrieve and parse page protection requests that can be linked to the page protection dataset by Hill and Shaw (2015) alongside ORES scoring of revisions, (ii) describing the characteristics of page protections and requests from January 2012 to March 2023, and (iii) conducting a quasi-experimental study to provide the first analysis of how page protections affect article quality.

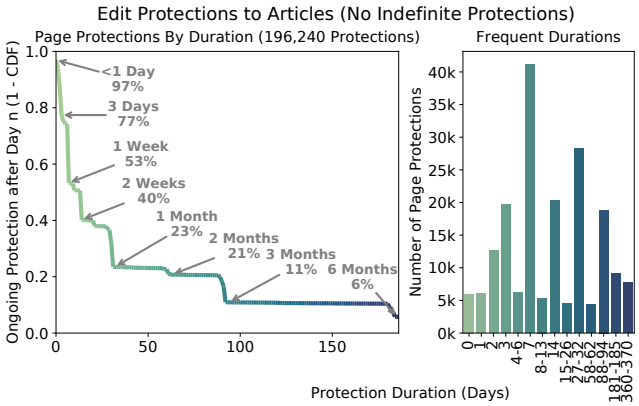


Figure 3: **Edit protections on the English Wikipedia.** We show the duration of 196,240 temporary article protections between January 2012 and March 2023. Most protections are enforced for a week, but three-day, two-week, one-month, and three-month protections also occur regularly.

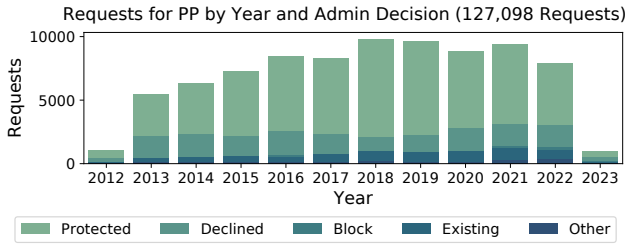


Figure 4: **Requests by year and decision.** Admins enforce protection in response to most requests (65.7%), while 20.2% are declined, 6.4% lead to user interventions (e.g., blocking), 6.3% are already protected, and 1.3% contain other responses (e.g., withdrawal of the RfPP).

4 Dataset

We now discuss the dataset and its preprocessing while also characterizing page protections. Our dagt are based on the openly available Wikipedia dumps, and all the resources (data, descriptive statistics, and code) required to reproduce the analyses in this paper are made openly available.[2] To avoid misuse of the dataset in terms of user anonymity, we do not publish any parsed usernames.

Page protection spells. We process page protection spells similarly to Hill and Shaw (2015). This approach only includes protections with at least the level of semi-protection and leaves out pending changes protections. We obtain 488,772 protection spells starting after December 31st 2011, and for this work, only consider edit protections (299,721), dropping page move or creation protections. Of these 299,721 spells, 215,746 regard article pages ("namespace 0"), while the remainder concern other pages, such as images or talk pages. We further remove spells for articles that were moved (i.e., name changes) during a spell (total of

[2]https://zenodo.org/records/13349480

2,502 protections). Of the remaining spells, 17,004 protections are active as of March 2023, with a median protection time of 1,003 days and 932 spells being upheld for less than 32 days. We exclude these indefinite protections in our final dataset of 196,240 protections (Fig. 3; Table 2).

These 196,240 spells were applied to a total of 104,305 unique pages—with the articles about Basketball player *DeMarcus Cousins* and the country *Turkey* being the most frequently protected (30 and 29 spells, resp.).[3] Meanwhile, 67,131 articles were only protected once. Overall, this signals a frequent rotation of temporary protections while certain articles have longer, indefinite protection. This aligns with Wikipedia guidelines, as they consider long or indefinite protections of pages harmful and state that such protections should only be applied where warranted (e.g., the main page) or in case of persisting disruptive behavior even after temporal protections expire (Wikipedia 2021, 2023a).

Page protection requests. We parse 163,227 requests for page protection (RfPP) from the English Wikipedia archives (Wikipedia 2022b) from October 2012 until March 2023. As the archives are structured similarly to Wikipedia talk pages, we adopt and extend existing code from MWChatter[4] and mwparserfromhell[5] to parse their wikitext (i.e., Wikipedia's "source code"). We discard requests (1) with malformed wikitext or article names, without a valid timestamp for the users' initial request, or without a final decision by an admin (3.7% of archived requests); (2) for pages not in the article namespace (21,520); (3) from actions other than editing (6,815); and (4) for pending changes protection (7,794). We then select the latest admin response to the resulting 127,098 valid RfPPs and group them into five groups (Table 2; Fig. 4): protected (83,487), declined (25,718), declined but carried out an intervention affecting certain users (e.g., user block or referral to the edit-warring noticeboard; 8,169), protection existing (8,096), or others (e.g., withdrawal of request; 1,628).

Merging requests and first protections. To exclude confounders that might arise due to multiple protections such as previous history of disruptive editing, we only include the first semi-protections of articles (94,298 spells) in our analysis of the effect of page protections on quality. We thus merge first protection spells that started after the implementation of the RfPP form in October 2012 (88,239 spells) with accepted requests for page protections. We find requests for 34,401 spells (39% of first protections), suggesting that most first protections are actually enforced by admins rather than requested by users. While we are unable to match around 4% of accepted RfPP to spells, and we previously noted that we could not parse around 3% of the original RfPPs from the archives, this still represents a considerable amount of non-requested page protections.

Article metadata. We utilize the Mediawiki history dataset (Wikimedia Foundation 2023) as well as the ORES API (Halfaker and Geiger 2020) to collect metadata for our articles. We use the dumps to gather revision data such as edit counts and article length and ORES to predict article quality and topic.

[3]D. Cousins: https://w.wiki/3ovx, Turkey: https://w.wiki/3hLW

[4]https://github.com/mediawiki-utilities/python-mwchatter

[5]https://github.com/earwig/mwparserfromhell

Edit Protections	**Temporary (Ended)**	**Matched Study**
Obs. period	01/2012 – 03/2023	01/2013 – 12/2022
Count	196,240	102,156
Requests for Page Protection	**Parsed from Archives**	**Matched Study**
Obs. period	08/2012 – 03/2023	01/2013 – 12/2022
Count	127,098	48,308
% Accepted	65.69	50.00
% Declined	20.23	50.00
% Others	14.08	00.00

Table 2: **Datasets.** We provide an outline of the full dataset as well as the matched dataset for our quasi-experimental study. Due to matching with replacement, declined RfPPs may occur multiple times in the matched dataset.

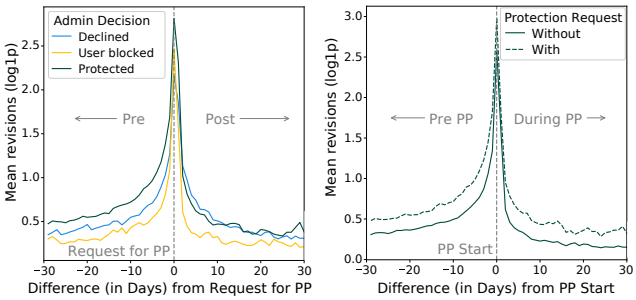
Listing 1: **Example of quality prediction and score.** We show an excerpt of the JSON generated by an API call to the ORES *articlequality* model for the April 2024 revision of the Wikipedia article "Association for the Advancement of Artificial Intelligence" (https://w.wiki/9sBDm). Based on the weighted sum of probabilities for all quality levels (Halfaker 2017), this article revision has a quality score of 1.438 (= $0.027{\cdot}0{+}0.634{\cdot}1{+}0.233{\cdot}2{+}0.089{\cdot}3{+}0.011{\cdot}4{+}0.005{\cdot}5$).

```
"score": {
  "prediction": "Start",
  "probability": {
    "Stub": 0.027, "Start": 0.634, "C": 0.233,
    "B": 0.089, "GA": 0.011, "FA": 0.005,
  }
},
"features": {
  "feature.wikitext.revision.chars": 4815.0,
  ...
  "feature.len(<datasource.english.idioms>)": 0.0
}
```

Article topic. We collect the predicted top-level topic labels based on ORES' *articletopic* taxonomy (*Culture*, *STEM*, *History and Society*, and *Geography*).[6] We only assign a topic to a revision if ORES predicts a likelihood for the topic > 0.5 for the majority of five revisions made before the request or start of protection, and drop all articles without a topic. Note that an article can have multiple top-level topics.

Article quality. For article quality, we use ORES' *articlequality* model, which predicts quality from structural features of articles (e.g., number of sections and references or article length). This quality assessment provided by ORES is a machine learning classifier that (i) was developed by (past) members of the Wikimedia Research Team, (ii) was trained and evaluated on Wikipedia data, (iii) was manually evaluated by editors and tool developers, (iv) can readily be computed for large datasets as a transparent metric through an API, (v) and in practice correlates with other interpretations of quality, such as writing or factual accuracy (MediaWiki 2023; Halfaker and Geiger 2020).

[6]ORES Articletopic endpoint (https://w.wiki/7Rrd)



(a) Requests for page protection. (b) Enforced page protections.

Figure 5: **Edits around page protection key events.** We observe that both for RfPP *(a)* as well as enforced protections *(b)*, spikes in activity are dampened by the corresponding event, and edits revert to levels prior to the intervention.

The ORES *articlequality* endpoint predicts a label based on the probability assigned to one of six quality classes present in the English Wikipedia: Stub, Start, C-, B-, Good, and Featured Articles (in ascending order). As research has demonstrated that categorical prediction of quality may be prone to mislabeling (Warncke-Wang, Cosley, and Riedl 2013) and a continuous quality metric allows for more precise assessment, we follow the approach by Halfaker (2017) to combine predictions made by ORES into a continuous quality score between 0 and 5. For this, we multiply the enumeration of all ordered quality classes (Stub = 0 to Featured Article = 5) by the corresponding prediction probability for each class and sum these scores. Listing 1 exemplifies the output of the ORES quality prediction API as well as the calculation of our weighted quality metric.

Editing patterns by admin decision and protection type. We analyze editing activity by notable admin decisions for first protections around the RfPP and the start of protection (Fig. 5). In particular, Fig. 5a suggests that RfPPs are preceded by edit spikes, followed by a decrease immediately after the request, regardless of an admin's decision. As the submitted RfPP notifies an administrator, they might deny protection but intervene in other ways, e.g., warning users and reverting individual conflicting revisions. Furthermore, the increased activity might fade away as users could naturally decrease their editing or lose interest in a topic. Second, there appear to be lower activity levels in articles that are protected without a prior RfPP compared to protections of articles with an RfPP (Fig. 5b). Therefore, RfPPs seem to generally affect editor activity irrespective of an admin's page protection decision.

5 Methods

We create a quasi-experimental study (see Hernán and Robins (2016)) that simulates the page protection process as the following randomized experiment: when admins receive an RfPP, they flip a coin and accept the RfPP if it lands on "heads" or reject the RfPP if it lands on "tails." Then, in the 13-week period following the decision, they track the article quality of protected and non-protected pages to determine the effect of protection on quality. Note that this experiment is not practically feasible, as RfPPs are not accepted or declined "at random." Therefore, we describe how we approximate this experiment using matching and difference-in-differences regression.

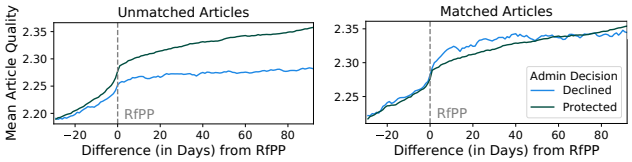


Figure 6: **Unmatched and matched mean article quality**. We note differences in trends in the unmatched data before the request for page protection (RfPP), which are not evident in the matched data, indicating good quality of the match.

Data. We select the observational period from January 2013 until December 2022, as the English Wikipedia implemented RfPP in late 2012, and the processed dumps extend until March 2023. We analyze articles that were not previously protected and disregard RfPP of articles that were protected or received another RfPP within 90 days to minimize potential confounding effects from prior protection phases or special administrative oversight. Additionally, we exclude first protections longer than 91 days. Our experiment's pre-treatment period spanned four weeks before up until the actual request, and we cover 13 weeks (91 days) in the follow-up period after the RfPP, where we consider an article to be consistently treated. Early investigations suggested that the end of protection does not result in significant changes in activity or article quality trends (Appendix, Fig. 12). We consider the RfPP as the time of treatment, as administrators can initiate clean-up activities immediately after a protection request is submitted, even before protection is formally applied. Nonetheless, protections are generally enforced close to the RfPP (median $=$ 117 minutes, mean $=$ 282.23 minutes, SD $=$ 478.32 minutes). Finally, we consider the last recorded quality per article and week as the target variable.

Matching. The data described above allows us to compare articles that had their protection requests approved with those that had not. However, a simple comparison may be problematic, as unlike in our hypothetical randomized experiment, in reality page protections are not assigned randomly. On the contrary, there may be confounders that can influence both the assignment of treatment and outcome. To account for these, we use the pre-treatment period to match protected ("treated") to non-protected articles ("control") using the following features measured before the RfPP:

- **Activity.** Edits 1 hour, 24 hours, and 1 week before $(\log(x+1))$.
- **Controversy.** Number of reverts (i.e., reversal of previous edits) 1 hour, 24 hours, and 1 week before $(\log(x+1))$, as a proxy of conflict (Sumi, Yasseri et al. 2011).
- **Article length.** Maximum article length in bytes (log).
- **Article age.** Overall number of edits to the article (log).
- **Quality.** Maximum article quality 1, 8, and 21 days before the intervention. We do not use finer-grained measures as potential content fluctuations might obscure actual quality.

- **Topic.** Top-level topics (History and Society, STEM, Culture, Geography). An article can have more than one topic.

We employ propensity score matching (Hill and Reiter 2006) but require exact matching on article topic(s) and use caliper matching for quality (caliper $= 0.25$, one-quarter of a quality level) as well as the propensity score (one standard deviation). After matching, absolute standardized mean differences (Stuart et al. 2011) for all covariates are below 0.1 (Appendix, Fig. 13). This approach produces 24,154 matched control-treatment pairs after discarding 1,401 rejected (control) and 1,463 accepted (treatment) RfPPs for which we did not find matches (Tab. 2). We plot the mean article quality for matched and unmatched articles in Fig. 6.

Difference-in-differences model. Comparing outcomes between treatment and control groups could uncover the effect of page protections, provided that the matching variables control for all causal paths between treatment and outcome (Pearl 2009; Appendix). Nonetheless, we use a difference-in-differences approach to make our results more robust. We compare treated (protected) with control articles (not protected) under the parallel trends assumption, i.e., that in the absence of treatment, the difference between the treatment and control group remains constant over time. Intuitively, this method creates a counterfactual estimate of how the treatment group would have progressed without treatment and compares it to the observed change in the outcome.

The difference-in-differences approach enables us to estimate the causal impact of the protection on the article quality using a simple linear model. We consider a panel of $a = 1, \ldots, N$ units for $t \in T$ relative periods and estimate coefficients δ_t using a fixed effect regression of the form

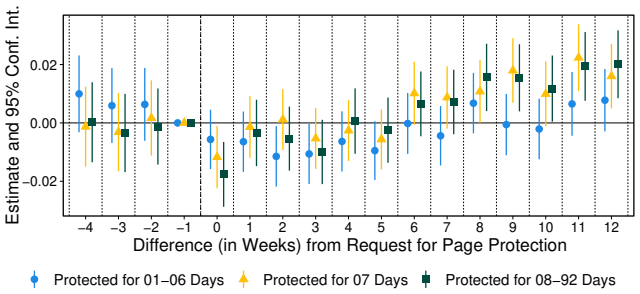
$$Y_{a,t} = \alpha_a + \sum_{t \in T} \delta_t D_t + \epsilon_{a,t}, \tag{1}$$

where $Y_{a,t}$ is the outcome associated with article a at time t, α_a are unit fixed effects, and D_t are indicator variables for unit a being t periods before or after treatment. We measure the effect of treatment on outcome in time t via δ_t and use the R package *fixest* for computation (Bergé 2018). Due to our matching, we assume that the parallel trends condition holds before the intervention, and we attempt to prevent anticipation for the treatment (protection) by setting our treatment to the time of the RfPP instead of the protection start. We do not account for time-fixed effects as we consider the relative date of the request for page protection.

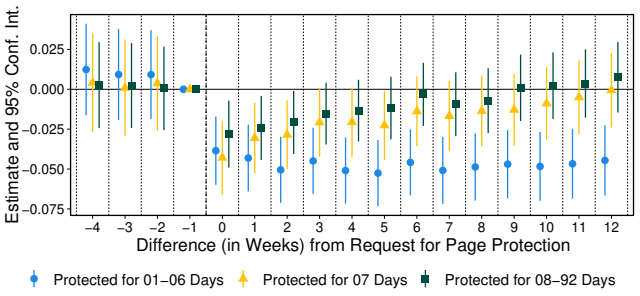
6 Results

We now present the measured effects of protection on quality. Note that we report results for z-score normalized values of our target variable (i.e., quality), meaning that an increase of 0.1 equals an increase by 10% of the standard deviation.

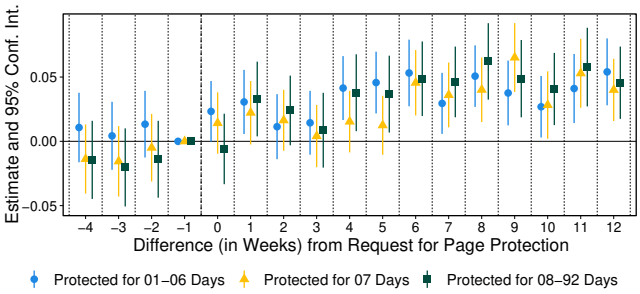
Short-term decrease, long-term recovery. In Fig. 7a, we visualize weekly effects for all protections lasting < 7 days (8,949 pairs), 7 days (8,401 pairs), and > 7 days (6,804 pairs). We observe that for protections lasting 7 or > 7 days, quality shows a minor but significant decrease in the week after the request (-0.017 and -0.018, $p < 0.05$, resp.),



(a) All articles.



(b) Low-quality articles.



(c) High-quality articles.

Figure 7: **Difference in article quality after protection.** We use difference-in-differences models to analyze effects on the matched dataset and in (a) find small differences in quality between articles with declined and accepted requests. In (b) and (c), we fit the same model on low-quality and high-quality articles, and detect heterogeneous effects.

while protections of less than 7 days observe no significant changes (-0.006, $p > 0.05$). In the weeks thereafter, as quality recovers to pre-request baselines around week 6, articles subject to protections for 7 days or longer even significantly raise their quality until week 12 (0.016 and 0.02, $p < 0.05$, resp.). Overall, we conclude that page protections in the short term seem to reduce quality, but in the long term, to increase quality. However, although reported effects are significant, the overall effect size is small (e.g., an increase of around 2% of the standard deviation in week 12).

Low-quality articles suffer, high-quality articles improve. We now categorize articles into low quality (lowest quartile, quality < 1.71; Fig. 7b) and high quality (highest quartile, quality > 2.94; Fig. 7c). Low-quality articles experience slight but significant decreases right after protec-

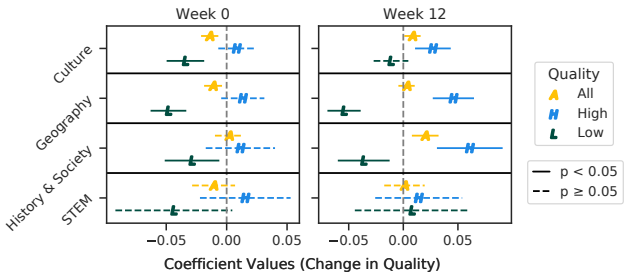


Figure 8: **Effect of page protection by topic.** Low-quality articles show a decrease right after protection (week 0) and mostly do not improve in the long term (week 12), while high-quality articles generally improve. Notably, STEM is an exception in both subsets, as it remains relatively stable.

tion (e.g., $-0.043, p < 0.05$ for 7-day protections). By week 12, the levels recover to baseline for most, except for those with less than 7 protection days, which sustain a level of quality similar to the one observed during the initial decline ($-0.045, p < 0.05$ in week 12). For high-quality articles, we found no significant changes in the week after the protection ($p > 0.05$ for all durations). Instead, these articles slightly improve their long-term quality (e.g., 7-day protections in week 12, $0.058, p > 0.05$).

Differential effects for topics. Topics strongly shape attention and content on Wikipedia (Kobayashi et al. 2021; Singer et al. 2017). Therefore, we combine all protection durations and investigate effects for Culture, Geography, History & Society, and STEM for all articles as well as only low-quality and high-quality articles (Fig. 8). We find mostly homogeneous effects for topics in low-quality articles right after protection, with small decreases and a more negative effect in the long term (e.g., Geography -0.048 for week 0, -0.062 for week 12). For various topics in high-quality articles, we observe improved quality over time, with most topics showing a gradual increase in quality (e.g., History and Society, $0.061, p > 0.05$ for week 12). Notably, effects are not significant for STEM articles ($p > 0.05$).

Article length, other factors also influence quality. Next, we examine an important component of article quality: length. Similar to our previous approach, we categorize articles into quartiles according to pre-RfPP length (log), dividing them into shorter (lowest quartile, < 9.96) and longer articles (highest quartile, > 10.34), and employ our previous model with length as the target variable (log scale, z-score standardization; Appendix, Fig. 15). Overall, the results are similar to what we found for quality. Articles that were already long prior to protection became even lengthier by week 12 ($0.07, p < 0.05$; Fig. 15), relatively to shorter ones.

However, we find that length does not fully explain the increased quality we observe. For all 26 article features employed by ORES' *articlequality* model, such as the number of sections, links, and references (MediaWiki 2023; Halfaker and Geiger 2020), we perform a variation of our difference-in-difference analysis, comparing *week -1* with *week 0* and *week 12*, respectively (features are log-scale, z-score standardized). We find that our overall quality is also driven by other features such as the number of references and links (Fig. 9). For example, after protection, low-quality articles contain more "citation needed" templates and have fewer references (week 0, $p > 0.05$). On the contrary, in the weeks following protection, high-quality articles significantly increase the number of headings, references, and links, among other things (week 12, $p < 0.05$).

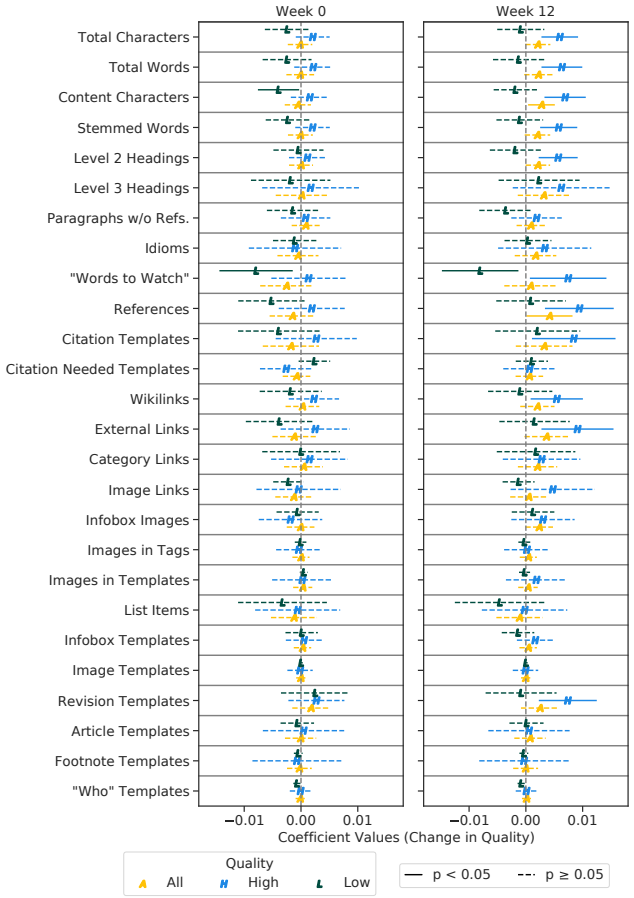


Figure 9: **Effects of individual features on quality.** Quality decreases and increases are driven by several factors included in the ORES model. For example, long-term comparison (week 12; right) to unprotected articles suggests that protected low-quality articles observe a decrease in "citation needed" templates and references, while protected high-quality articles increase their length and number of structural features, such as citations and links.

Robustness of the analyses. We varied our matching configuration and difference-in-differences model to ensure robustness. First, we placed narrower (0.1) and wider (0.5) calipers on the quality prior to the request. Next, we performed exact matching on the year and month of the request. Additionally, we fit a daily version of the model along with the weekly one. We did not observe any considerable changes to our main findings in any of the variations.

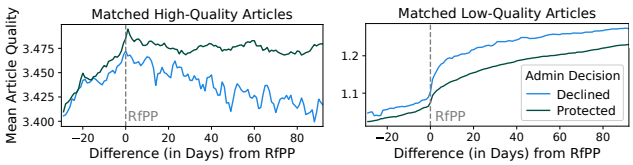


Figure 10: **Article quality in low- and high-quality articles by protection status.** Right after their request for page protection (RfPP), protected low-quality articles display a slightly greater improvement in mean article quality compared to their unprotected counterparts (right), while unprotected high-quality articles undergo a decline in quality after the RfPP and protected ones maintain their quality (left).

7 Discussion

We now discuss our findings and their implications, limitations, and highlight future avenues for research.

Low- and high-quality articles differ in effects. We find that page protections have a positive effect on high-quality articles but a negative effect on low-quality ones. To better understand the mechanism driving this differential effect, we further analyze the mean article quality in our matched dataset, splitting by quality level and protection status (Fig. 10). This reveals an important observation: protected high-quality articles display resilience in maintaining their quality levels after the intervention, while non-protected high-quality articles exhibit a consistent decline in quality. Conversely, trends for low-quality articles indicate that unprotected articles experience a bigger increase in quality after protection, compared to protected articles.

These patterns could be due to low-quality articles having greater potential for improvement when unprotected, as even small changes can be valuable to the content. In that context, page protection reduces these articles' already limited chance of improvement, as on Wikipedia, a vast mass of articles are competing for attention at any time (Ribeiro et al. 2021). Furthermore, possible vandalism reverts or misinformation corrections by experienced editors or admins after protection could result in collateral damage due to excessive content removal. This issue may affect shorter, low-quality articles more heavily than longer, high-quality ones. While high-quality articles can usually withstand the removal of a sentence or two, shorter ones may be disproportionately impacted. In contrast, protected high-quality articles fare better than their unprotected counterparts, as without protection, new contributions may be of lower quality (e.g., fewer references) or vandalize the text. For these better articles, protection can even be viewed as a structured approach for quality improvement (Warncke-Wang et al. 2015), which, for example, is triggered when articles compete for promotion to the featured article of the day.[7]

Not all topics are volatile. The development of protected articles relating to History and Society, Culture, or Geography also exhibits this differential effect. Across these topics, protections appear to have a detrimental impact on low-quality articles, but a positive effect on high-quality articles. In contrast, STEM articles seem less prone to quality changes. This could be attributed to the fact that STEM articles usually comprise more universal truths based on the nature of their subject matter, resulting in more consistent content. Conversely, topics such as "Culture" or "History & Society" frequently involve articles about subjects susceptible to change and controversy, such as biographies of living persons and articles about wars or conflicts. These findings complement past work by Yasseri and Kertész (2013), which suggests that edit wars are more common in articles related to religion, politics, and geography.

Information quality and ORES article quality. Information quality is a difficult-to-assess metric. This is especially relevant for Wikipedia, considering its millions of articles across many language editions (Moas and Lopes 2023). Therefore, as elaborated in our methodology, we employ predictions made by the well-established ORES *articlequality* endpoint (Halfaker and Geiger 2020)—which predicts quality based on structural features of article texts—to interpolate our quality metric (Halfaker 2017; Listing 1). While this quality assessment does not directly assess knowledge integrity (Aragón and Sáez-Trumper 2021), factual correctness, or writing quality, the basic structural article features considered for ORES' quality prediction are nonetheless associated with these more advanced article characteristics. For example, ORES takes into account references, citations, and templates (Piccardi et al. 2020; Redi et al. 2019), which serve as proxies of other notions of information quality.

Ramifications for editors. Besides directly affecting content, page protection's administrative barriers can hamper the growth of Wikipedia by discouraging editors, as previous work has shown that a similar negative impact on engagement occurs when edits by new users are reverted (Halfaker et al. 2013). This is not typically a concern for high-quality articles, as they often cover popular subjects or receive frequent surges in attention that draw new users and content. However, it may pose a serious problem for niche topics that receive little contribution. Although past research has suggested that protection does not necessarily drive away editors (Ajmani, Vincent, and Chancellor 2023; Klapper and Reitzig 2018), its effects may vary between novice and experienced users, analogous to the variations observed in low- and high-quality articles. Further examining the effect of protections on editors is a promising avenue for future work.

Admin decisions and their motivation. Wikipedia admins act as judges in the case of article protection. As studies have shown the influence of biases and ideology on the decisions made by judges in a court of law (Harris and Sen 2019), such factors may also affect the actions of Wikipedia admins (Das, Lavoie, and Magdon-Ismail 2016) and therefore merit further investigation.

Moreover, in "Characterizing Page Protections", we were only able to find RfPPs for less than half of our edit protections. This observation is significant, as it suggests reduced community participation in the protection process and frequent admin decision-making without editor input. Protection of articles without a previous RfPPs could result from

[7] Wikipedia:Today's featured article: https://w.wiki/5TW$

spillover activity (Zhu, Walker, and Muchnik 2020) from other protected articles, for which a RfPP may or may not have been submitted. Alternatively, administrators may protect these articles based solely on their own judgments (Das, Lavoie, and Magdon-Ismail 2016), because of expanded authority in case of community sanctions, or after being notified of conflicts through "watching" or "patrolling" articles.[8] In combination with the findings of this study, future work should aim at uncovering these particular protections.

Practical implications. Our research has practical implications for page protection guidelines on Wikipedia. In addition to the existing policies (Wikipedia 2023a), admins should consider an article's characteristics before deciding on protection. The findings of our study indicate that protecting high-quality articles carries minimal risk. However, greater caution seems necessary when protecting low-quality articles, particularly on specific topics.

Limitations. We utilize the well-established ORES article quality predictor as our primary source of quality measurement and establish robustness by extending our experiments to article length, which has been shown to strongly correlate with the general measurement of information quality on Wikipedia (Blumenstock 2008). To further supplement this, future research could conduct interviews of casual readers and experienced Wikipedians to qualitatively investigate the differences between protected and unprotected content.

Furthermore, our scope is limited to the English Wikipedia. While protection levels, RfPP archives, quality levels, and their predictors vary across language, our experiments, code, and data can still provide a framework and serve as a baseline for the analyses of other Wikipedia versions. As community dynamics and policies vary across languages (e.g., the Portuguese Wikipedia prohibits anonymous edits entirely), the effects of page protection on article quality could also differ (Johnson and Lescak 2022).

8 Conclusion

Page protection is a core policy of Wikipedia, enforced by administrators to prevent content from harm by limiting contributions. Although this mechanism protects the online encyclopedia from "evil," such as vandalism, it can also prevent "good" by hindering article development. In this work, we aim to assess how page protections on the English Wikipedia affect article quality, as measured by structural features of articles. We characterize protections and user requests for protection for over a decade of data and provide, to the best of our knowledge, the first quantitative assessment of the effect of page protection on quality. Using a quasi-experimental study, we find differential effects: As high-quality articles benefit, low-quality ones may diminish in quality following protection. Furthermore, effects vary across different topics. These findings indicate that protecting high-quality articles may present a low risk, but protecting low-quality articles on certain topics requires caution. Overall, our findings shed light on one of Wikipedia's most important content moderation tools and inform improvements to administrative processes on the Web.

[8] See WP:General_sanctions (https://w.wiki/AjAh), WP:Watch (https://w.wiki/4W33), and WP:Patrol (https://w.wiki/7kDy).

References

Ajmani, L.; Vincent, N.; and Chancellor, S. 2023. Peer Produced Friction: How Page Protection on Wikipedia Affects Editor Engagement and Concentration. *CHI*, 7(CSCW2).

Aragón, P.; and Sáez-Trumper, D. 2021. A preliminary approach to knowledge integrity risk assessment in Wikipedia projects. *arXiv preprint arXiv:2106.15940*.

Bergé, L. 2018. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13).

Blumenstock, J. E. 2008. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, 1095–1096.

Burke, M.; and Kraut, R. 2008. Mopping up: modeling wikipedia promotion decisions. In *CSCW'08*, 27–36.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *HCI*, 1(CSCW): 1–22.

Das, S.; Lavoie, A.; and Magdon-Ismail, M. 2016. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *TWEB*, 10(4): 1–25.

DeDeo, S. 2016. Conflict and computation on Wikipedia: A finite-state machine analysis of editor interactions. *Future Internet*, 8(3): 31.

Faulkner, R.; Walling, S.; and Pinchuk, M. 2012. Etiquette in wikipedia: Weening new editors into productive ones. In *WikiSym*, 1–4.

Halfaker, A. 2017. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In *OpenSym*, 1–9.

Halfaker, A.; and Geiger, R. S. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *HCI*, 4(CSCW2): 1–37.

Halfaker, A.; Geiger, R. S.; Morgan, J. T.; and Riedl, J. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5): 664–688.

Harris, A. P.; and Sen, M. 2019. Bias and judging. *Annual Review of Political Science*, 22: 241–259.

Hernán, M. A.; and Robins, J. M. 2016. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8): 758–764.

Hickman, M. G.; Pasad, V.; Sanghavi, H. K.; Thebault-Spieker, J.; and Lee, S. W. 2021. Understanding wikipedia practices through hindi, urdu, and english takes on an evolving regional conflict. *HCI*, 5(CSCW): 1–31.

Hill, B. M.; and Shaw, A. 2015. Page protection: another missing dimension of wikipedia research. In *WikiSym*, 1–4.

Hill, J.; and Reiter, J. 2006. Interval estimation for treatment effects using propensity score matching. *Statistics in medicine*, 25(13): 2230–2256.

Horta Ribeiro, M.; Cheng, J.; and West, R. 2022. Post Approvals in Online Communities. *ICWSM*, 16(1): 335–346.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *HCI*, 5(CSCW2): 1–24.

Johnson, I.; and Lescak, E. 2022. Considerations for multilingual wikipedia research. *arXiv:2204.02483*.

Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007. He says, she says: conflict and coordination in Wikipedia. In *CHI*, 453–462.

Klapper, H.; and Reitzig, M. 2018. On the effects of authority on peer motivation: Learning from Wikipedia. *Strategic management journal*, 39(8): 2178–2203.

Kobayashi, R.; Gildersleve, P.; Uno, T.; and Lambiotte, R. 2021. Modeling collective anticipation and response on Wikipedia. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 315–326.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *TheWebConf*, 857–866.

Lemmerich, F.; Sáez-Trumper, D.; West, R.; and Zia, L. 2019. Why the world reads Wikipedia: Beyond English speakers. *WSDM'19*, 618–626.

Lorini, V.; Rando, J.; Saez-Trumper, D.; and Castillo, C. 2020. Uneven coverage of natural disasters in Wikipedia: The case of flood. *arXiv preprint arXiv:2001.08810*.

MediaWiki. 2023. ORES: Article quality. https://w.wiki/7g4C. Accessed: July 13, 2023.

Moas, P. M.; and Lopes, C. T. 2023. Automatic Quality Assessment of Wikipedia Articles—A Systematic Literature Review. *Computing Surveys*, 56(4): 1–37.

Pearl, J. 2009. *Causality*. Cambridge university press.

Piccardi, T.; Redi, M.; Colavizza, G.; and West, R. 2020. Quantifying engagement with citations on Wikipedia. In *Proceedings of The Web Conference 2020*, 2365–2376.

Redi, M.; Fetahu, B.; Morgan, J.; and Taraborelli, D. 2019. Citation needed: A taxonomy and algorithmic assessment of Wikipedia's verifiability. In *The World Wide Web Conference*, 1567–1578.

Ribeiro, M. H.; Gligorić, K.; Peyrard, M.; Lemmerich, F.; Strohmaier, M.; and West, R. 2021. Sudden attention shifts on wikipedia during the covid-19 crisis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 208–219.

Ruprechter, T.; Horta Ribeiro, M.; Santos, T.; Lemmerich, F.; Strohmaier, M.; West, R.; and Helic, D. 2021. Volunteer contributions to Wikipedia increased during COVID-19 mobility restrictions. *Scientific reports*, 11(1): 21505.

Shi, F.; Teplitskiy, M.; Duede, E.; and Evans, J. A. 2019. The wisdom of polarized crowds. *Nature human behaviour*, 3(4): 329–336.

Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why we read Wikipedia. In *TheWebConf*, 1591–1600.

Spezzano, F.; Suyehira, K.; and Gundala, L. A. 2019. Detecting pages to protect in Wikipedia across multiple languages. *Social Network Analysis and Mining*, 9: 1–16.

Stuart, E. A.; King, G.; Imai, K.; and Ho, D. 2011. MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of statistical software*.

Suh, B.; Convertino, G.; Chi, E. H.; and Pirolli, P. 2009. The singularity is not near: slowing growth of Wikipedia. In *WikiSym*, 1–10.

Sumi, R.; Yasseri, T.; et al. 2011. Edit wars in Wikipedia. In *International Conference on Social Computing*, 724–727. IEEE.

Vaseva, L.; and Müller-Birn, C. 2020. You shall not publish: Edit filters on English Wikipedia. In *OpenSym*, 1–10.

Viégas, F. B.; Wattenberg, M.; Kriss, J.; and Van Ham, F. 2007. Talk before you type: Coordination in Wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, 78–78. IEEE.

Warncke-Wang, M.; Ayukaev, V. R.; Hecht, B.; and Terveen, L. G. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 743–756.

Warncke-Wang, M.; Cosley, D.; and Riedl, J. 2013. Tell me more: an actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, 1–10.

Wikimedia Foundation. 2023. Analytics: MediaWiki History. https://w.wiki/7kTP. Accessed: June 13, 2023.

Wikimedia Foundation. 2024. Wikimedia REST API. https://w.wiki/J8K. Accessed: August 08, 2024.

Wikipedia. 2010. User:ClueBot_NG. https://w.wiki/7bZJ. Accessed: June 13, 2023.

Wikipedia. 2021. Page protections considered harmful. https://w.wiki/7Ges. Accessed: June 13, 2023.

Wikipedia. 2022a. Wikipedia:Requests for page protection. https://w.wiki/39wf. Accessed: June 13, 2023.

Wikipedia. 2022b. Wikipedia:Requests for page protection/Archive. https://w.wiki/6paY. Accessed: June 13, 2023.

Wikipedia. 2023a. Wikipedia:Protection policy. https://w.wiki/6pac. Accessed: June 13, 2023.

Wikipedia. 2023b. Wikipedia:Wikipedians. https://w.wiki/4KuN. Accessed: June 13, 2023.

Yasseri, T.; and Kertész, J. 2013. Value production in a collaborative environment: sociophysical studies of Wikipedia. *Journal of Statistical Physics*, 151(3-4): 414–439.

Zheng, L.; Albano, C. M.; Vora, N. M.; Mai, F.; and Nickerson, J. V. 2019. The roles bots play in Wikipedia. *HCI*, 3(CSCW): 1–20.

Zhu, K.; Walker, D.; and Muchnik, L. 2020. Content growth and attention contagion in information networks: Addressing information poverty on Wikipedia. *Information Systems Research*, 31(2): 491–509.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, as Wikipedia is a public good that largely profits the general public.
   (b) Do your main claims in the abstract and 2 accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Characterizing Page Protections and Quasi-Experimental Setup.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Characterizing Page Protections.
   (e) Did you describe the limitations of your work? Yes, see Discussion.
   (f) Did you discuss any potential negative societal impacts of your work? We believe that advancing knowledge about Wikipedia in our case has no foreseeable negative societal impacts.
   (g) Did you discuss any potential misuse of your work? We discussed potential misuse of usernames in Characterizing Page Protections, which is why we do not publish any user information.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? See previous answer and Characterizing Page Protections.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, we discussed the potential positive and negative effects of page protection, e.g., in the Introduction.
   (b) Have you provided justifications for all theoretical results? Yes, e.g., in Quasi-Experimental Setup.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see Introduction, Quasi-Experimental Setup, and Discussion.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see Discussion.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, see Discussion and Limitations.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, at least the literature regarding Wikipedia.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see Discussion.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in Quasi-Experimental Setup.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? We do not run any machine learning experiments besides simple linear regression that can be run on any state-of-the-art CPU.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we clearly lay out the limitations of our experiment and how we aim to address our confounders.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in our case with a strong focus on the effect of possible confounders in our differences-in-differences setup.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? Wikipedia content are available under the CC-BY-SA and the GFDL, the Mediawiki history dumps under CC0, and code by Hill and Shaw (2015) under GNU GPL 3.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we include code in the GitHub and Zenodo repository.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? We do not gather any personal data about individuals.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? We exclude usernames from our data. Revisions containing offensive content are deleted from the Wikimedia datasets by default.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? We follow typical principles as other Wikimedia datasets.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? We did not create a separate Datasheet for the accompanying data for this paper. We however describe the most important details for data composition, collection, and preprocessing alongside corresponding statistics within this text.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A Appendix

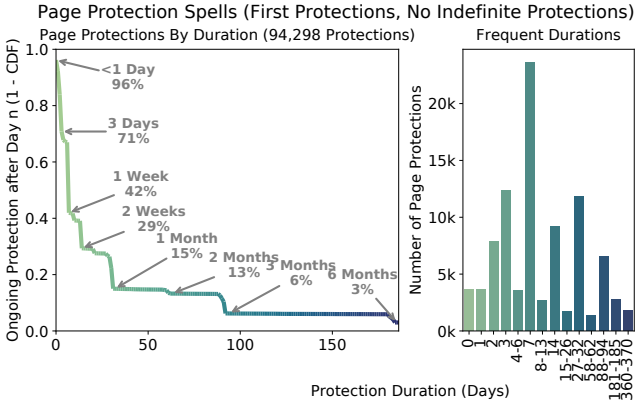


Figure 11: **First edit protections on the English Wikipedia.** Duration distribution of only the 94,298 first article protections does not differ considerably from our full dataset (Fig. 3). Most protections are enforced for one week, but three-day, two-week, one-month, and three-month protections also occur regularly.

Data

Missing data. We retrieve RfPPs starting from October 2012. However, due to an error in Wikipedia's archiving, October 2013 is missing from the archives and our dataset.

First protections. We show statistics about first protections in Fig. 11 (48% of all protections). Protection duration does not vary considerably between first and all protections.

Editing activity after protection ends. We illustrate edits prior to and following protection in Fig. 12 and observe that lifting page protection does not considerably increase edits.

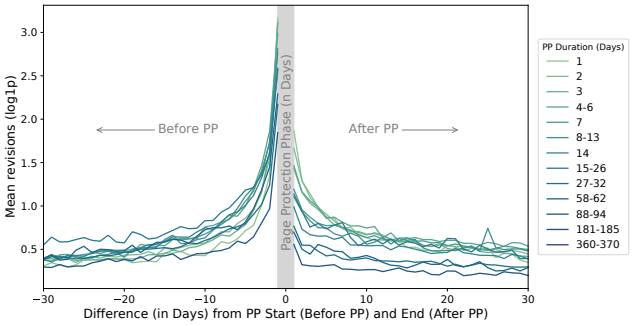


Figure 12: **Page protections by duration.** We find that editing patterns suggest no substantial decreases after the end of protections compared to baseline activity prior to the spikes that precede protection, regardless of protection duration.

Methods

More on treatment and control groups. We compare articles with an accepted protection request to those that were declined protection (Section "Quasi-experimental study"). We accept that the act of requesting article protection (even in the case of a declined request) can already be considered a treatment, as it possibly attracts the attention of admins and other experienced editors, who may perform minor janitorial actions on the page (e.g., selected reverts). Note that we thus do not have a "true control" in our experiment, but instead use a "quasi control" in the form of declined RfPPs. The comparison of a treatment of interest (i.e., protection) with other "baseline treatments" (i.e., declined RfPPs) is considered valid in designs where there is no true randomized control group (Hernán and Robins 2016). As in our case, activity patterns on non-contentious articles may not be comparable to the increased attention that pages subject to a RfPP receive and might introduce additional confounders.

Dataset balance. To assess the quality of our matches, we compute the absolute standardized mean differences for our full as well as our matched dataset (Stuart et al. 2011). We find that matching improved the covariate balance in our dataset, as absolute standardized mean differences for all covariates are below 0.1 for the matched dataset (Fig. 13).

Results

We perform multiple sensitivity analyses for our findings.

Article length. We categorize articles into quartiles according to pre-RfPP length (log), dividing them into shorter (lowest quartile, < 9.96) and longer articles (highest quartile, > 10.34). We then fit our model with article length as the target variable (log scale, z-score standardization) and find similar trends as for quality, as articles that were already long prior to protection became lengthier by week 12, relatively to shorter ones (Fig. 15).

Individual ORES features for matching. We extend the feature set for computing the propensity score for our article matching approach by also including the individual 26 ORES article quality features (min-max normalized). After recomputing the matches including these variables, we find

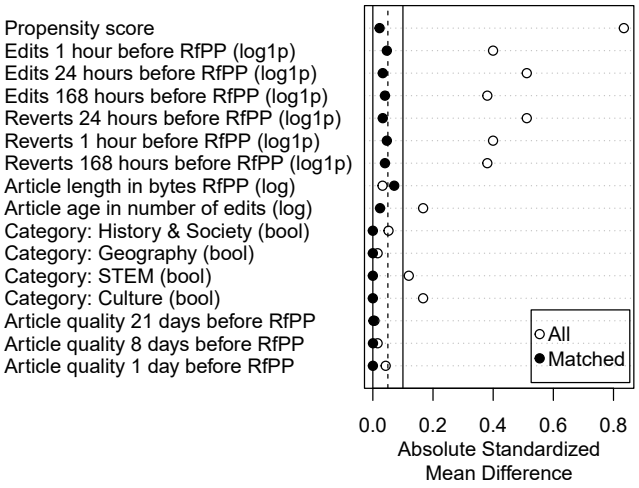
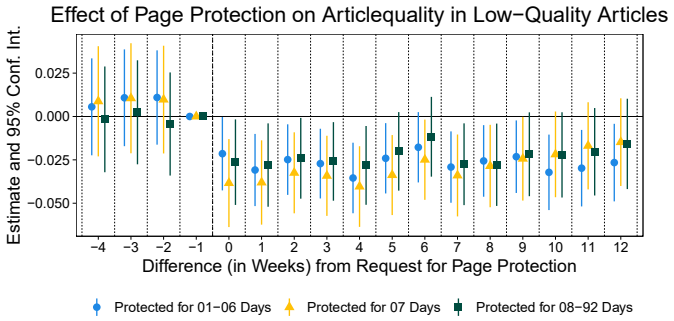


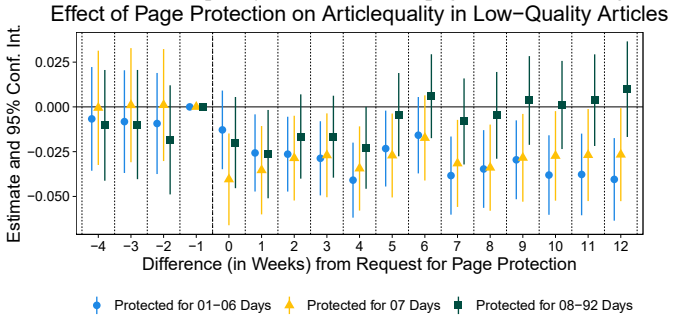
Figure 13: **Dataset balance after matching.** We show the absolute standardized mean differences before and after matching for all covariates in a "Love" plot. The dashed (at 0.05) and solid lines (at 0.1) signal that matching considerably improved covariate balance in the matched dataset.

no considerate difference in the trends of the results reported in the main section (Fig. 16).

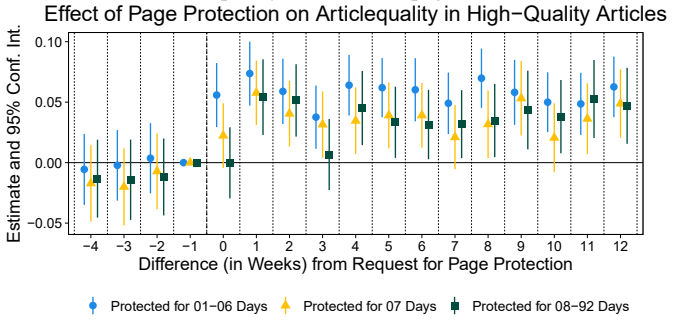
Attention (page views). Article attention or the number of page views might affect the behavior of the Wikipedia community towards protected pages and thus affect article quality, even though in practice views generally correlate strongly with edits (Hickman et al. 2021), which we already account for in our original matching procedure. Nonetheless, we extend our analysis with a robustness experiment that incorporates views as a measure of pre-RfPP attention. We utilize the Wikimedia API to retrieve page views (Wikimedia Foundation 2024) the day and week before the request and add these features to the propensity score computation for our article matching approach before again fitting a difference-in-differences regression. However, the Wikimedia API only provides page view data from July 2016 onward. As we report our main results for 2013 to 2022, the results of this sensitivity analysis now pertain to a different dataset—only the articles with an RfPPbetween July 2016 and 2022. Fig. 14 presents a comparison between results for high- and low-quality articles for two types of experiments within this time period: One with our previous matching procedure without page views (Figs. 14a and 14c) and one with a matching approach that includes page views (Figs. 14b and 14d). Altogether, the results depicted in these figures confirm that trends for high and low-quality articles mostly persist regardless of whether page views are incorporated into the matching process.



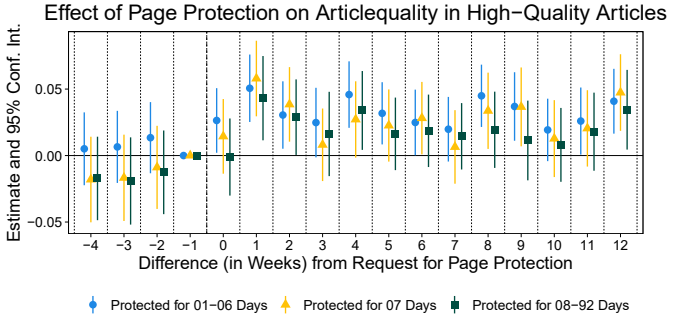
(a) Effects for low-quality articles without page views matching.



(b) Effects for low-quality articles with page views matching.

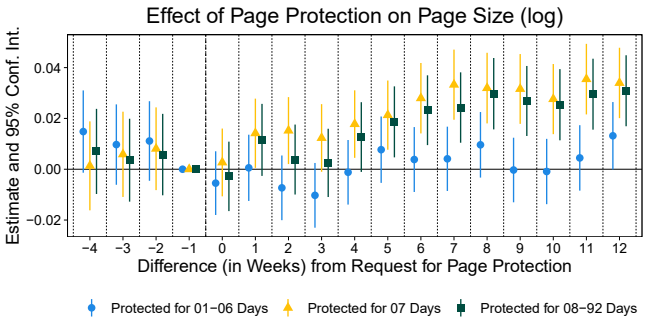


(c) Effects for high-quality articles without page views matching.

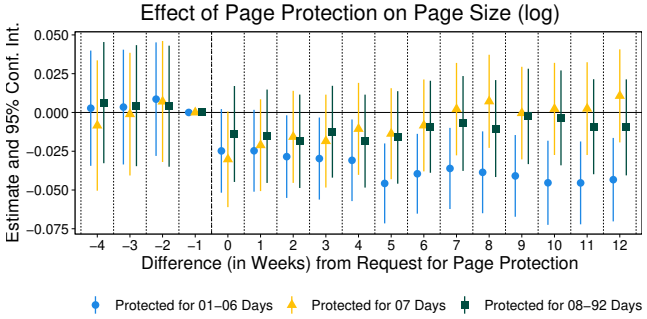


(d) Effects for high-quality articles with page views matching.

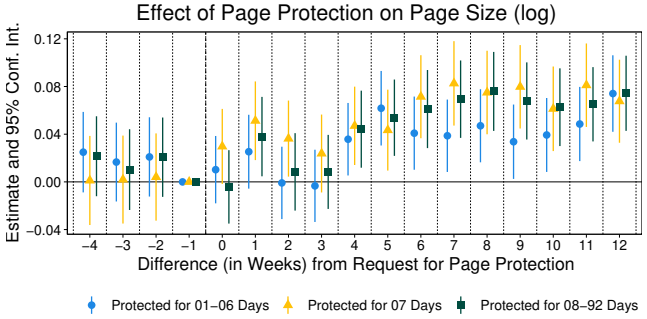
Figure 14: **Difference in article quality when considering page views.** As page views are only available after July 2016 through the Wikimedia API, we fit our dynamic difference-in-differences model on this reduced dataset (July 2016 to 2022) to compare results obtained without (left) and with (right) page views incorporated in the matching procedure.



(a) Effects for all Articles.

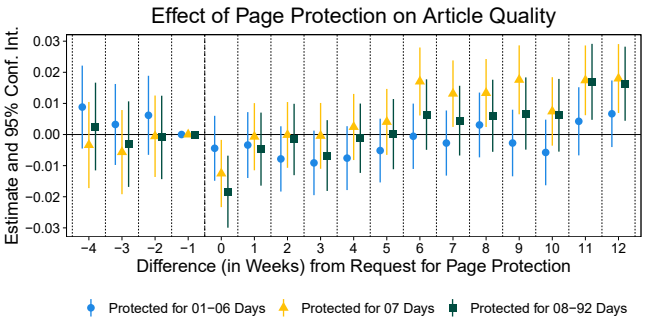


(b) Effects for shorter articles.

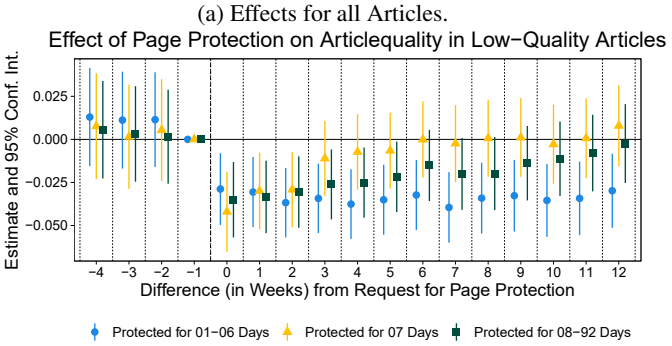


(c) Effects for longer articles.

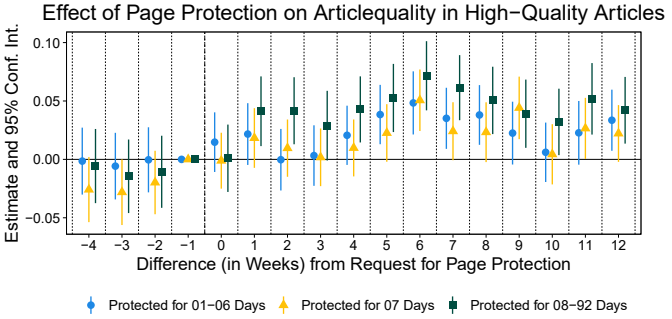
Figure 15: **Difference in article length after protection.** Results in (a) indicate only minor differences in length between declined RfPP and protected articles right after the request, but length rises to levels significantly higher than articles without protection. In (b) and (c), we detect heterogeneous effects for shorter and longer articles.



(a) Effects for all Articles.



(b) Effects for low-quality articles.



(c) Effects for high-quality articles.

Figure 16: **Difference in article quality when accounting for all ORES features.** We incorporate all 26 individual features utilized by ORES (see Fig. 9) in the matching procedure before again fitting our dynamic difference-in-differences regression. This sensitivity analysis does not significantly change our results.