The risks, benefits, and consequences of pre-publication moderation: Evidence from 17 Wikipedia language editions

ANONYMOUS AUTHOR(S)

Many online communities ~~and peer production communities~~ rely on on post-publication moderation ~~of incoming contributions~~. Contributors—even those that are perceived as being risky—are allowed to publish material ~~openly and~~ immediately. ~~Only after being made public is material checked, reviewed, and moderated~~. An alternative arrangement involves moderating all content before publication. A range of communities have argued against pre-publication ~~review~~ by suggesting that it makes contributing less enjoyable for new ~~users~~ and that it will distract established community members with ~~unnecessary~~ moderation work. We present an empirical analysis of the effects of a pre-publication review system called *FlaggedRevs* that was deployed by ~~numerous~~ Wikipedia language editions. We ~~collected~~ panel data from eighteen large wikis and test a series of hypotheses related to the effect of the system on indicators of activity levels and ~~average~~ quality within the affected communities. While there is some evidence that the system discouraged participation of unregistered users ~~after taking down their substandard contributions~~, ~~it did not make an impact on new and returning~~ registered editors. Our findings imply that concerns about the negative effects of pre-publication moderation systems on the quality, productivity, and sustainability of communities may be ~~small~~.

ACM Reference Format:
Anonymous Author(s). 2021. The risks, benefits, and consequences of pre-publication moderation: Evidence from 17 Wikipedia language editions. 1, 1 (April 2021), 42 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Successful commons-based peer production communities struggle to maintain the quality of the knowledge bases they construct against rising tides of vandalism, trolls, and spam [26]. ~~More~~ Recent work has shown that moderation systems can result in enormous collateral damage to communities. For example, rising vandalism in English Wikipedia led to increased rates of newcomer rejection and, ultimately, to a decreased contributor base [20, 21, 28, 43]. This dynamic appears to occur in a range of similar peer production

Manuscript submitted to ACM

contexts [22, 48]. Despite the costs, some form of content moderation is necessary to help enforce community or platform policies by monitoring and nullifying damaging contributions [17, 29, 51]. An important question for social computing scholars and designers asks which types of moderation systems can provide strong protection while minimizing harm to communities.

One important distinction between types of moderation systems used in social computing systems reflects when moderation occurs relative to publication of user-contributed information. Many platforms use *post-publication moderation* where content contributed by users is made instantly visible online and remains there until it is changed or removed by other users or automated systems. Although less common, *pre-publication moderation* involves attempting to prevent malicious contributions proactively. While this brings the obvious benefit of mitigating threats of vandalism, these systems can require a large operational cost to effectively assess all contents submitted in a timely manner. As a result, pre-publication review systems are typically applied only to a subset of contributions from individuals perceived as less trusted or higher risk by community administrators (e.g., users contributing without accounts or new users).

That said, the choice to deploy pre-publication moderation systems in communities remains controversial with some communities adopting it and many others rejecting it. For example, many of the largest communities on Wikipedia (including English) have

discussed the use of these systems at length but ultimately rejected them.[1] While there is some agreement that additional quality assurance mechanisms can be helpful in fighting vandalism, there is deep concern about collateral damage from these systems. How much attempted vandalism will a shift from post-publication to pre-publication moderation deter? How many vandals will be able to work around the system? How many good contributions will never occur in a pre-publication moderation context that would have happened in a post-publication moderation context?

In this paper, we present a quantitative evaluation of a pre-publication review system called *FlaggedRevs* and its deployment on the 17 Wikipedia communities. Using data available from these wikis, we used a community-level panel data *interrupted times series analysis (ITS)* analysis as well as a user-level *general linear mixed model (GLMM)* to identify the effects of FlaggedRevs on several different outcomes. Measured in a variety of ways, we find that the introduction of pre-publication moderation has a surprisingly minor impact. While the system successfully prevented many low quality contributions from ever being visible to the public, it did not appear to affect the volume or quality of contributions overall, contributions made by users with accounts, or the return rate of newly registered contributors. Although the system caused a decline in the productivity of unregistered users, we cannot rule out the idea that at least some of these users may

[1] https://meta.wikimedia.org/wiki/Talk:Flagged_Revisions

simply have created accounts in order to build trust within the new system. Our analysis supports the argument that some level of pre-publication moderation for contributions from high-risk users may be a useful approach for communities seeking to maintain order ~~while minimizing collateral damage~~.

~~This~~ work makes several contributions to the social computing literature. First, we contribute to theories of peer production by formally articulating a set of novel theoretical claims related to the effects of pre-publication moderation systems based on prior work and conversations in communities. Second, our work makes a series of empirical contributions by testing these claims as hypotheses. ~~Our work makes empirical contribution to the study of peer production systems and Wikipedia by evaluating the FlaggedRevs~~ pre-publication review system in 17 different Wikipedia communities. ~~Ultimately,~~ we argue that, contrary to expectations and most of our hypotheses, the system ~~does not~~ raise transaction costs sufficiently to inhibit participation by the community as a whole nor ~~does it~~ measurably improve the quality of the wikis on which it is implemented.

2 BACKGROUND

Peer production describes a widespread and influential type of online collaborative ~~production~~ that involves the mass aggregation of small contributions from diversely motivated individuals working together over the Internet [5]. Although the most famous

examples of peer production include free/libre open source software and Wikipedia, the model of production also describes collaborative filtering on sites like Reddit, Q&A on sites like Quora, and the activity on a range of other knowledge bases. Theorized first in 2002 by Benkler [3, 4], peer production is ~~characterized by~~ extremely low transaction costs ~~made possible by advances in new communication technology~~. In other words, peer production is possible because it extremely easy to contribute to relative to existing models of organizing production, ~~e.g.~~ markets and firms. ~~This allows contributions from people who are often only slightly~~ motivated to participate. ~~Because it is often very easy to contribute, communities can attract large numbers of of contributions from a diverse sense of perspectives [4, 6].~~ The low barrier to entry ~~also~~ helps stimulate participation across both ~~groups of~~ unregistered and registered users [2], and introduces ~~more~~ diverse perspectives from a wider pool of participants [1, 23]. Classical theories of public goods and collective action ~~have shown~~ that low participation costs help communities achieve critical mass [32], facilitate fluid or open community boundaries [37], and attract greater diversity of perspectives [4, 6, 23].

Prior research on peer production communities has shown that malicious actions such as trolling, spamming, and flaming can undermine a community's purpose and drive members away [12, 26, 30, 31]. To minimize these harmful behaviors, communities rely on teams of volunteer and professional moderators who review content submitted by

~~users and review~~ low quality or norm-violating ~~material~~ [41, 42]. When the presence of harmful ~~or repetitive~~ content is negated, a healthy environment can encourage positive social norms and continued participation from existing contributors and motivate new participants to join [34]. Indeed, a study by Wise et al. shows that "a moderated online community elicited greater intent to participate than an unmoderated community" [51].

In peer production sites, most moderation ~~is~~ *post-publication* ~~moderation~~ in that it occurs after content has been submitted and made fully public. Post-publication moderation is often considered ~~as~~ preferable from an user experience perspective for several reasons. First, it is relatively low-cost because not all content must be monitored. It is also potentially efficient because more attention will naturally be given to more popular content. Finally, it is thought to be effective at eliciting contributions because it allows for real-time engagement ~~for everyone~~ leading to faster collaboration and an immediate sense of belonging and self-efficacy in contributors who see their contribution made "live" immediately. When ~~edits~~ are immediately reflected and new information can be updated instantly, it incentivizes writers to continue contributing to keep the information updated ~~in a timely manner~~ [50]. ~~On the other hand~~, post-publication moderation is ~~entirely reactive and is always~~ at least partially ineffective in fighting vandalism ~~attempts damage is always done before it is taken down~~—even if only briefly.

~~However, as communities grow, content governance can become more labor intensive and complicated. Higher volumes of contributions naturally come with higher volume of low-quality contributions, whether intentional or not.~~ The reactive nature of post-publication moderation systems means that malicious behaviors such as trolling, flaming and spamming can potentially go unnoticed and lead to low satisfaction among users [8]. ~~In order to maintain the quality and efficiency of the community~~, additional measures are sometimes considered to weed out malicious behavior. *Pre-publication moderation* is an alternative approach where contributors must wait for their work to be explicitly approved by a moderator before becoming visible to the general public. Because this review work ~~by experienced or trusted community members~~ can be ~~onerous,~~ this moderation approach is frequently focused on monitoring only certain groups of users that have not yet earned the trust of the community ~~due to an insufficient contribution history~~. ~~These would be the group of~~ unregistered users (i.e., those who contribute to the site without registering for an account ~~with a stable identification~~) ~~or~~ newly registered users.

For example, Quora—a popular crowdsourced Q&A site—began requiring contributions from unregistered users to be reviewed and approved before being published in 2017.[2] Although English Wikipedia uses post-publication review almost exclusively [15], German Wikipedia—the second largest version in terms of article ~~volumes~~ and active contributors

[2] https://techcrunch.com/2017/02/10/qa-site-quora-clamps-down-on-anonymity-will-review-content-before-publishing-restricts-actions/

of the encyclopedia—designed and deployed a pre-publication moderation system called *Flagged Revisions* (more commonly known as *FlaggedRevs*) to require that contributions made by unregistered or new and "untrusted" registered users to be reviewed and approved (i.e., "flagged") before they were visible in the version of the article shown to most viewers. The FlaggedRevs system has subsequently been deployed in twenty-three other wikis on Wikipedia.

Because pre-publication review is often only used to review contributions from relatively high-risk or untrusted users, its effects are likely uneven across users. In particular, we separately hypothesize its effect across three groups of users: (a) affected users whose work is subject to review (typically new or unregistered users), (b) users whose work is not subject to review (typically established users with a history of contributing) who will often be involved in reviewing the contributions from the former group, and (c) the community overall.

Pre-publication review is designed to minimize damage caused by bad actors and in turn discourage antisocial behaviors by encourage positive examples and enforcing standard social norms [44, 46]. We assess whether the system is indeed functioning as intended by offering a single hypothesis in three parts: *pre-publication moderation will be associated with a smaller number of low quality contributions from from affected users that are made visible to the public (H1a)*. Because the work of unaffected users are not

subject to review, we expect that *pre-publication moderation will not be associated with a a difference in the number of low quality contributions from from unaffected users that are made visible to the public (H1b)* (i.e., we expect a null effect). Further, we also expect to see an overall reduction in the number of substandard contributions are made public. In other words, *pre-publication moderation will be associated with a smaller number of low quality contributions overall that are made visible to the public (H1c).*

In a related sense, we anticipate that *pre-publication moderation will be associated with a higher contribution quality from the less experienced and untrusted group (H2a).* Once again, we also expect a null effect for the unaffected user group, stating that *pre-publication moderation will not significantly affect the quality of contributions from established users (H2b).* Finally, we also expect that *pre-publication moderation will be associated with higher contribution quality of the community overall (H2c).*

Next, we turn to research that has shown that additional quality control policies may negatively affect the growth of a peer production community [20, 21]. For example, a higher barrier to entry is likely to deter participation[35, 36]. Collateral damage in the form of deterred or rejected good faith or quality contributions is difficult to measure but if frequently cited by community members as a concern with moderation systems in general, and with pre-publication review systems in particular.

A range of past social computing studies suggests that additional quality control policies tend to negatively affect the growth of a peer production community, as new users are particularly sensitive to the feedback they receive for their good-faith, but low quality, contributions [13, 20–22]. Our third set of hypotheses anticipates that the benefits described in H1 and H2 come with a trade-off related to the types of collateral damage shown in the evaluation of other moderation systems. Because of the reduced sense of efficacy associated with contributing, we anticipate that *pre-publication moderation systems will be associated with reduced contributions from affected users (H3a).*

Although the intuition is less clear, we also believe that previous work also suggests that a pre-publication system will also negatively effect the contribution rate of established users whose contribution are not directly subjected to review. There are several possible explanations for this scenario. First, more experienced contributors are required to review content so the deployment of the system may increase the demands on these users' times and compete for their time. Additionally, the system may discourage members of affected groups from returning to the site to eventually become veteran members. As a result, we hypothesize that *pre-publication moderation systems will be associated with reduced contributions from members of unaffected users (H3b).* Since both our previous hypotheses point to reduced contribution rates, we also suggest that *pre-publication moderation*

systems will be associated with reduced contributions from members of the community overall (H3c).

Finally, we considered that contributors of user-generated content sites often seek rewards from their peers' recognition of their work, either by feedback or status [10, 11]. Reducing this immediate reward from contributors might diminish their enthusiasm for long-term commitment [18]. Attracting and retaining new users is very important for peer production communities to maintain and expand their growth [21], and the higher barrier to entry, combined with delayed intrinsic reward, might be disheartening enough to drive these newcomers away. Different from our previous hypotheses, our fourth hypothesis focuses solely on new users. We anticipate that *the deployment of a pre-publication moderation system would negatively affect the return rate of newcomers (H4).*

Spoiler alert: with the exception of H1, we found null effects for most of our hypotheses, suggesting that the theories underpinning our hypothesis development might require revisiting.

3 EMPIRICAL SETTING

As one of of the largest peer production websites on the Internet, Wikipedia describes itself as "the free encyclopedia that anyone can edit"[3]. Wikipedia allows editing by the general public with very little in the way of additional barriers. Most visitors can edit the large majority of Wikipedia pages without creating an account, providing an email address, or establishing a persistent identity. Unfortunately, the ease of editing not only attracts legitimate editors to participate, but also can allows malicious and bad-faith participants to undermine the effort to maintain a high quality of content [45]. Wikipedia relies on transparency of all content updates to defend against the insertion of problematic information: every edit that is made is recorded, including the state of the article before and after the edit, and available for public review. Revisions found to violate policy may be removed and the state of the article easily reverted to the previous version. However, this approach of open editing privileges, transparency, and the revert to previous versions carries some risks. While larger communities may be able to field sufficient volunteer resources to fight vandalism and build, customize, and test advanced automated anti-vandalism tools [16?], volunteers in smaller communities may not have the time to expend on this effort. Automated tools to help protect the 321 different language editions of Wikipedia are an active area of research but the task is far from

[3] https://en.wikipedia.org

complete [19]. Uncaught vandalism, hoaxes, and disinformation could remain hidden for extended amount of time and detrimentally affect the community's reputation and its sustainability [27]. Consequences of uncaught vandalism are not unique to Wikipedia – for example, when the Los Angeles Times site experimented with a new online feature that "allowed readers to rewrite an editorial on the newspaper's website", they cancelled it soon after, due to rampant vandalism attempts by users.[4]

In an attempt to fix the aforementioned problem, the Wikimedia Foundation and the German Wikimedia chapter (Wikimedia Deutschland) have collaborated to develop *Flagged Revisions*,[5] (or *FlaggedRevs* for short), a highly configurable extension to the MediaWiki software that runs Wikipedia. FlaggedRevs acts as a pre-publication content moderation because FlaggedRevs software will display the most recent "flagged" revision of any page for which FlaggedRevs is enabled instead of the most recent revision in general. FlaggedRevs is designed to "give additional information about quality", by ensuring the readers of Wikipedia that all *flagged* revisions are vetted for vandalism or substandard content (i.e. contains obvious mistakes because of sloppy editing) and to provide information to readers about whether or not readers are reading a version of the article that has been reviewed by an established contributor.[6]

[4] https://www.latimes.com/archives/la-xpm-2005-jun-21-na-wiki21-story.html
[5] https://meta.wikimedia.org/wiki/Flagged_Revisions
[6] https://meta.wikimedia.org/wiki/Flagged_Revisions

Although there are many details that can vary based on the way that the system is configured, it has typically been deployed in the following way on Wikipedia language editions. Users are divided into groups of trusted and untrusted users. Untrusted users typically include all users without accounts as well as users who have created accounts recently and/or contributed very little. While unregistered editors remain untrusted, editors with accounts are automatically promoted to trusted status when they clear certain thresholds dictated by the wiki's configuration. For example, German Wikipedia allows registered editors who have made at least 300 edits with at least 30 edit comments.[7]

Revisions to articles made by trusted users are automatically made public (i.e., "flagged")—just as it would if FlaggedRevs were not deployed. Revisions to articles made by untrusted users are not made visible immediately but are instead marked as provisional and placed into a queue for review by others. These contributions must be reviewed by some other group (frequently, but not necessarily, the group of trusted users described above) who can flag (i.e., publish) the proposed revision, reject the proposed revisions by reverting it, or edit the proposed revision. The FlaggedRevs extension must be installed on a per-wiki basis meaning that it exists in some Wikipedia language editions, but not in most. Implementations of FlaggedRevs were placed under a moratorium in April 2017 due to the high staffing cost associated with configuring and maintaining the system.[8]

[7] https://noc.wikimedia.org/conf/highlight.php?file=flaggedrevs.php
[8] ibid.

Prior to this moratorium, twenty-four language editions of Wikipedia enabled FlaggedRevs.[9] Additionally, some language editions of Wikipedia-related projects (Wiktionary, Wikibooks, Wikinews, Wikiquote, Wikisource) have also implemented FlaggedRevs.

Despite it's importance and widespread deployment, very little is known about the effectiveness of the system and its impact on the affected communities. A report made by members of the Wikimedia Foundation in 2008 gave a brief overview of the extension, its capabilities and deployment status at the time, but acknowledged that "it is not yet fully understood what the impact of the implementation of FlaggedRevs has been on the number of contributions by new users".[10] Our work seeks to address this empirical gap.

4 METHODS

4.1 Data

We first collected an exhaustive dataset of revision activities, as well as the content moderation activities that occurred on all Wikipedia language editions that enabled FlaggedRevs. This included 24 wikis in total. The datasets that we collected are publicly available in the Wikimedia Downloads website.[11] For each wiki, we downloaded the stub-meta-history XML dumps which contains the metadata for all revisions made to pages that are public (i.e., non-deleted) at the time that the database dump was created. We also

[9] Although it did not use the system, English Wikipedia implemented a similar vetting system called *Pending Changes*, but on a much smaller scale.
[10] https://meta.wikimedia.org/wiki/FlaggedRevs_Report_December_2008
[11] https://dumps.wikimedia.org/

used the flaggedrevs SQL dump which contains information about all the revisions that have been reviewed under the FlaggedRevs system in each wiki. ~~The datasets include~~ substantial variation across the wikis included including linguistic diversity, activity level, and organizational structure. Wikis vary in size, with numbers of contributions ranging from thousands to millions each year. Furthermore, the way each wiki configured FlaggedRevs also varies ~~enormously~~. Differences include the criteria for a revision to be flagged, published guidelines for determining the quality ~~and intent~~ of ~~the~~ revision, or differences in when the system will treat a contributors as "trusted." We used a series of custom software scripts to convert the data in both ~~the~~ database dumps into a tabular format which allowed us to build measures associated with the concepts in our hypotheses. We have placed the full code and datasets used for these analyses into an archival repository in the Harvard Dataverse (URL ~~not included~~ for blind review).

Because of the way we constructed our hypotheses, we built two datasets with different metadata and units of analysis. The first dataset is ~~referred to as the~~ *wiki-level dataset*. We use ~~this datasets~~ to test H1, H2, and H3. ~~It contains the aggregated information of all Wikipedia language editions that enabled FlaggedRevs, which is described further below~~. The second dataset is ~~referred to as the~~ user-level dataset, which is used to test H4. Each row of this second dataset represents information of an individual user ~~within~~ one of the wikis in our analysis.

In the wiki-level dataset, our unit of analysis is the wiki month. In each case, we constructed these data by aggregating across the raw dataset collected initially where each row involves a revision to a page in Wikipedia. We then proceeded to first restrict our analysis to only article pages by excluding revisions to pages in non-article "namespaces" (i.e., discussion pages, special pages, etc.).[12] We did so because Wikipedia has a different set of guidelines when reviewing these contributions for vandalism and FlaggedRevs may not enabled for these pages. Next, we aggregate data and looking at the total number of contributions by month for each wiki and further grouping them by type of user in ways that correspond to our sub-hypotheses (as described in our measures section below). Since our first three hypotheses concern the effect of the content moderation mechanism on active communities of affected and unaffected users, we excluded any wiki with fewer than 30 new contributions per month on average that are made by each of these editor groups.

For each published contribution, we must know the timestamp of the contribution, the timestamp of the it being published (i.e., flagged), and whether or not it is is published manually (as opposed to automatically and immediately). Because they are critical for our analyses, we omitted any wiki for which we could not obtain these data. As a result, our empirical setting is a population of 17 Wikipedia communities operating under different

[12] https://en.wikipedia.org/wiki/Wikipedia:Namespace

languages: Albanian, Arabic, Belarusian, Bengali, Bosnian, Esperanto, Persian, Finnish, Georgian, German, Hungarian, Indonesian, Interlingua, Macedonian, Polish, Russian and Turkish.

Because each wiki enabled FlaggedRevs at a different point in time, and we are only interested in observing possible immediate impact of the intervention, our datasets are restricted to revisions made to the 12 months periods before and after the day FlaggedRevs is enabled. Finally, because each wiki varies vastly is size, and we are only interested in the average effects of FlaggedRevs across all wikis, we standardized each measure of outcome in H1, H2, and H3 in standard deviation unit within individual wikis. For example, instead of averaging the number of contributions per month made by an editor group across 17 wikis, we calculated the number of contributions of each wiki, in *standard deviation unit*, before calculating the average number across all wikis. This ensures that the measure of outcome from each wiki has a equally weighted impact on our analysis, regardless of their size.

The user-level dataset does not have additional exclusion criteria but only contains information about new editors from the 17 wikis listed above. We consider each unique editor ID (either a pseudonym if it is a registered account, or an IP address) to be a unique user, and looking to the record of contributions of said user. Because each row in the dataset contains information of an editor, the total number of data points across

17 wikis would be so much bigger than the previous dataset. As a result, the dataset is only restricted to users whose contribution is made within 90 days before and after FlaggedRevs is enabled. There are 1,972,861 observations in this second dataset.

4.2 Outcome measures

Our H1 hypotheses pertain to the number of low quality contributions that are made visible to the public. We operationalize this as the *number of visible rejected contributions* which reflects the aggregated and standardized number of visible reverted contributions per month, for each wiki. On Wikipedia, when a contribution to an article is rejected, the moderator usually performs a "revert" action to nullify any changes said contribution has made. With FlaggedRevs, edits are never "disapproved" explicitly. Instead, rejected edits are simply reverted to a previously accepted version. Before FlaggedRevs is enabled on a wiki, all contributions are instantly accessible by the public, even if they are later reverted. After FlaggedRevs is enabled, contributions made by the affected editor groups would have to wait and be approved (flagged) to be visible.

Our H2 hypotheses suggest that pre-publication review will affect the quality of contributions overall, not only the ones that are made visible. We operationalize H2 in two ways. First, we use the *number of rejected contributions* which we operationize as the number of identity reverts. Because reverting is the most frequently used tool to fight

vandalism[24], the number of ~~rejected~~ contributions is a fairly reliable indication of quality. A large body of previous research into Wikipedia has used identity reverts as a measure of quality [24, 49?]. ~~This measure of outcome is also aggregated and standardized.~~ We also test ~~this~~ hypothesis using average quality which we operationalize as *revert rate*. Following previous research [47, 49], revert rate is measured as the proportion of contributions that are eventually reverted. Together with the number of ~~rejected~~ contributions, ~~they~~ give a more complete information about the quality ~~of work from each editor group~~.

Our H3 hypotheses call for a measure of user productivity. ~~Once again,~~ there a range of ways to measure productivity. We follow Hill and Shaw [22] and a range of other scholars who operationalize wiki-level productivity as the *number of contributions* (i.e., unique revisions or edits made) to article pages (i.e., pages in the article namespace). This means that we exclude things like conversation, governance, and interpersonal communication ~~that~~ are also contributions but ~~which~~ are made to non-article pages. ~~This measure is also aggregated by month and standardized by wiki.~~

Finally, we test our H4 hypotheses by looking for changes in *return rate* ~~which is~~ measured at the level of each user. First, we follow previous work by Geiger et al. to break edits into sessions which define as "a sequence of edits made by an editor where the difference between the time at which any two sequential edits are saved is less than

one hour" [14]. A user is said to have returned if they make their first edit session to a article page and then make another edit session within 60 days of their first edit session.

Our hypotheses are framed in terms of groups of users who are affected (H*N*a), not affected (H*N*b), and overall community effects (H*N*c). In practice, this means that we stratify each of the variables described above into ~~several~~ groups of user. We operationalize these groups as follows:

(1) **IP editors**: Editors who edit on Wikipedia without a registered ~~pseudonym~~ and whose contribution is credited to their IP address. This group's contributions are subject to FlaggedRevs so this measure is used to test ~~part of~~ our hypotheses H*N*a.

(2) **First-time registered editors**: Users who registered ~~for~~ an account and ~~made~~ their first edit. This group's contributions are also affected ~~to~~ FlaggedRevs so this measure is used in a second set of tests of hypotheses H*N*a.

(3) **Returning registered editors**: Users who had contributed at least one edit session under a stable identifier. Because each Wikipedia language edition makes choices to determine whether a returning registered editors is "trusted", and because these configurations have been changed over time, it is therefore extremely difficult to determine exactly whether a returning registered editor's work was automatically flagged. That said, it is safe to assume that a large number of contributions made by returning registered editors are not affected by FlaggedRevs because a large majority

of contributions to wikis belong to a very small group of veteran contributors [38]. Because the measure is imperfect, it acts as a conservative test of our hypotheses H*N*b.

(4) **All editors**: All users contributing to a wiki. This group is used to test our hypotheses H*N*c.

4.3 Analytic Plan

For our first three hypotheses, we seek to understand the impact of FlaggedRevs at the time it was implemented. To do so, we use Interrupted time-series analysis (ITS). ITS is a quasi-experimental research design and is particularly useful when investigators have neither control over the implementation of an intervention nor the power to create treatment group and control group [25]. ITS analysis involves constructing a time series of population-level outcome of interest (i.e., each of the measures described above) and then testing for statistically different changes in the outcome in the time periods before and time periods after implementation of the intervention [40]. ITS has been used in a series of social computing studies [9, 39, 46].

ITS relies on a series of assumptions and analytical choices [7]. First, it requires a clearly defined intervention with a known implementation date. In our dataset, each wiki has a clear date where FlaggedRevs is enabled. The second step is identifying an outcome

that is likely to be affected by the intervention. For each of our hypotheses, we clearly describe the affected measures as described above. It is an important reminder that, in our dataset, different wikis have vastly different sizes, meaning comparing the effect of FlaggedRevs across different wikis requires some data scaling. As a result, instead of using the raw number of contributions per time interval, we standardized the outcome of interest in standard deviation units. Third, ITS requires at least 8 observations of outcome at 8 different points in time for each period before and after the intervention. Usually, more observations are useful, but having too many data points might have diminishing effects, as there are cases when the measure of outcome changes significantly in the immediate period after the intervention, but ends up balanced out in the long run, which could affect the model. We chose to observe our independent variable on a monthly basis, over the period of two years, from 12 months before to 12 months after FlaggedRevs is enabled.

In practice, ITS divides the dataset into two time segments—the first segment comprises rates of the event before the intervention or policy, and the second segment is the rates after the intervention. ITS applies what is effectively a segmented regression to allow the researcher to test differences the change in level (i.e., a change in the intercept) as well as change in slope associated with the intervention or change in policy while controlling for the overall trend in the outcome rate of interest. Segmented regression essentially

means that a linear regression model is applied twice and there are separate intercept and slope coefficients for the pre-intervention and post-intervention time segments. In this way, ITS allows researchers to estimate a statistical difference between the two time periods ~~as a basis to verify the causal claim that the intervention has a meaningful impact on the outcome of interest.~~

ITS typically is conducted using a single time series. In our case, we have panel data from 17 different wikis—each with their own trajectories. As a result, we fit wiki-level baseline trends and then seek to estimate the average changes (in both intercept and slope) at the point in time that FlaggedRevs is enabled. The panel data linear regression model that ~~ITS uses~~ has the following form:

$$Y = \beta_0 + \beta_{\boldsymbol{wiki}}\,\boldsymbol{wiki} \times time + \beta_1 flaggedrev\_on + \beta_2 flaggedrev\_on \times time + \varepsilon$$

The descriptions of our variables in our model are as follows:

(1) Y: Our dependent variable capturing ~~volume of activity~~ from some subset of users, as described in the previous section.

(2) *time*: The month in the study period relative to when FlaggedRevs is turned on. For example, if we measure the outcome of interest one month prior to the day FlaggedRevs is implemented, the time variable would have a value equal to -1.

(3) *flaggedrev_on*: A dichotomous variable indicating the pre-intervention period (coded 0) or the post-intervention period (coded 1).

(4) **wiki**: A categorical variable included as a vector of dummy variables indicating the wiki (bold notation indicates that this variable is vector).

We do not report or interpret the coefficients associated with $\beta_{\boldsymbol{wiki}}$ which are included only as control variables and which capture baseline trends for each wiki (e.g., some might be increasing or decreasing in some measure). Because we have standardized each wiki in terms of activity, we do not need to include wiki-level intercept terms since these will all take the same value as a result of our measure construction. Our estimate for β_1 associated with *flaggedrev_on* estimates the average instantaneous change in level immediately following the intervention that enabled the system. Finally, the coefficient β_2 associated with (*flaggedrev_on* × *time*) indicates average change in slope following the intervention.

To test H4, we seek to understand how FlaggedRevs was associated with the return rate of new users. Because the measure of return rate happens at the level of individual users, we can do better than an overall aggregate and instead model the effect at the return of users. For this, we use a user-level dataset were the unit of analysis is a new user on each wiki. While it is possible that editors can contribute under different IP addresses, or creating different accounts, we assume that the first edit made by under

an unique editor id counts as a new user (they can be unregistered or registered). Our analytical model is a general linear mixed model (GLMM), based on the work by Mulloch et. al. [33]. In particular, we use a logistic binomial regression model with two wiki-level random effects at the wiki level: a random intercept term (1|**wiki**) and a random slope term (*time*|**wiki**). These address issues of repeated measures of users within wikis. The mixed-effects model we use is as follows:

$$\log(\frac{returned}{1-returned}) = \beta_0 + \beta_1 flaggedrevs\_on + \beta_2 first\_edit\_published + \beta_3 unregistered + \beta_4 time + \beta_5 first\_edit\_published \times unregistered + \beta_6 flaggedrevs\_on \times first\_edit\_published + \beta_7 flaggedrevs\_on \times first\_edit\_published \times unregistered + (1|\textbf{wiki}) + (time|\textbf{wiki})$$

The descriptions of our variables in our model are as follows:

(1) *returned*: a dichotomous variable indicating whether or not an editor made another edit session within 60 days of the first edit session.

(2) *flaggedrevs_on*: A dichotomous variable indicating whether or not an editor's first edit session was made after the day FlaggedRevs is implemented

(3) *first_edit_published*: A dichotomous variable indicating whether or not the final edit in a user's first edit session is published. If the entire session is reverted—as is common—this will revert all of the edits in the session including the final edit.

Because revisions can still be reverted after they are published, we measured that the editor's first edit session is visible if it is either not reverted at all, or it is reviewed (flagged) before being reverted.

(4) *unregistered*: a dichotomous variable indicating whether or not this is an IP editor.

(5) *time*: the month in the study period relative to when FlaggedRevs is turned on. For example, if we measure the outcome of interest one month prior to the day FlaggedRevs is implemented, the time variable would have a value equal to -1.

(6) **wiki**: a categorical variable of wiki-level as random effects, to fit individual trend lines for each wiki.

Our test for H4 is is focused on the parameters associated with the interaction terms between our three independent variables (*flaggedrevs_on*, *first_edit_published*, *unregistered*). These interactions allows us to understand how the effect of FlaggedRevs on newcomer return rate varies based on whether the user had registered for an account when making the first edit, and whether their contribution was visible to the public at some point. For example, we suspected that accepting or reverting someone's work as a form of feedback would likely have an impact on the contributor's decision to make further commitment to the site, especially among unregistered newcomers, an observation that was in line with the work of Halfaker et. al. [21]. As usual, our main focal point here is the effect of FlaggedRevs, and whether or not it has a statistical significance on newcomer retention

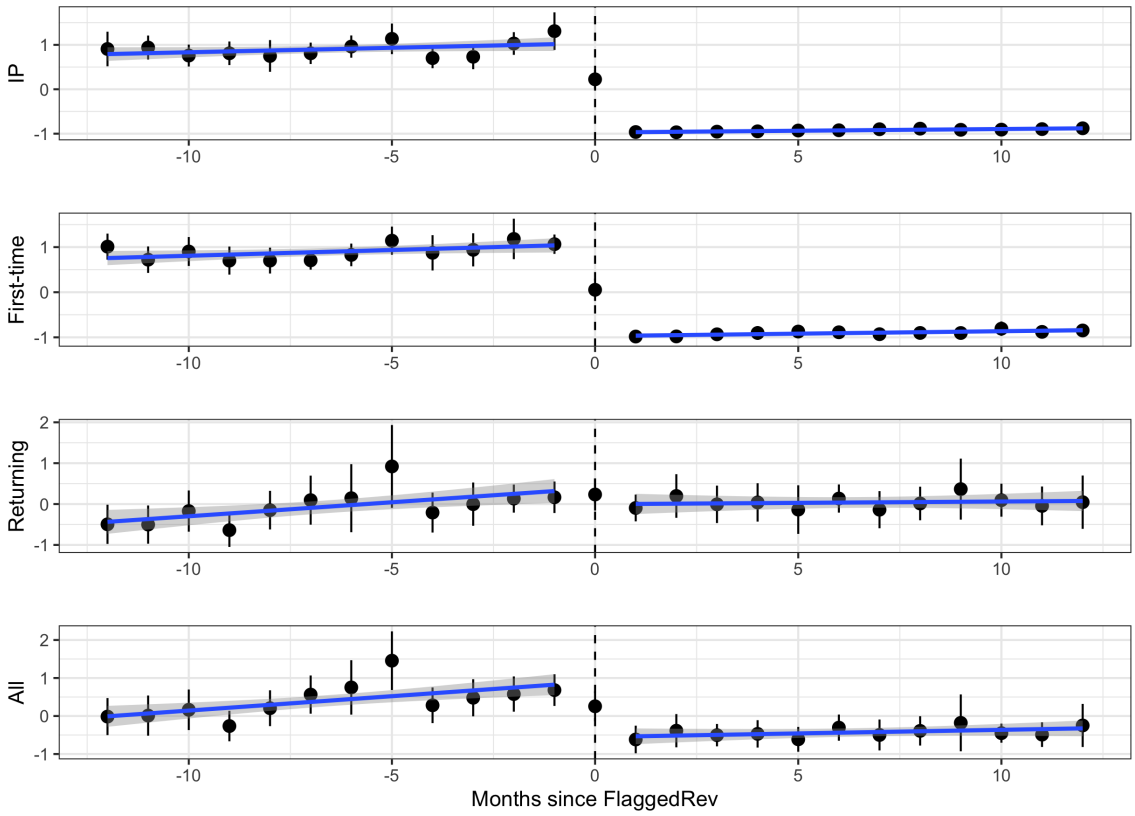

Fig. 1. Impact of FlaggedRevs on the quantity of visible reverted contributions (in std unit) made by different groups of users on Wikipedia. The vertical line indicates the start of the intervention. The regression lines along with confidence bands are fitted by a simple linear model. The error bars represent 95% confidence intervals.

~~rate. The *time* variable is another independent variable, which helps us understand the trend of the outcome through time.~~

5 RESULTS

To evaluate each of our hypotheses, we first look at a series of visualizations which allow us to visually inspect our data for the hypothesized relationships. Figure 1 shows the data we use to test the first set of hypotheses (H1) and depicts the ~~edit~~ trend of

our four user groups in two separate periods: pre-intervention and post-intervention. The vertical line indicates the start of the intervention. Each data point represents the mean of the number of contributions by an editor group, in standard deviation units. The regression lines along with confidence bands are fitted by a simple linear model (we use *geom_smooth()* function in R to draw these lines). The error bars represent 95% confidence intervals. We chose to clearly differentiate the two periods by excluding the data point at the month of the intervention. The table of coefficients and their result of statistical statistical significance for our ITS model will be presented in the Appendix, as we are going over the main findings.

From both the Figure 1 and our statistical test that is reported in Table ??, we see a very clear effect of FlaggedRevs in reducing the number of visible reverted contributions among affected users (H1a) as well as the overall community (H1c). It is clear that after the intervention, most of contributions made by the affected users that will eventually be rejected are rejected before they are published. Our statistical tests from our ITS analysis shown in Table ?? in the Appendix confirm that these effects are statistically significant. As we hypothesized, this relationship was not statistically significant among returning registered users who appear to be largely unaffected by the moderation system. The results we see are in line with our expectation for the first hypothesis and confirm that

FlaggedRevs is effective at preventing substandard contributions by affected users from being visible to the public.

Our tests for H2 are visualized in Figures 2 and 3 which allow us to visually assess the impact of FlaggedRevs on the volume of low quality contributions and average quality among the four user groups. At the month that FlaggedRevs is installed, we see the number of rejected edits made by IP editors spiked significantly. Upon further investigation, we find that when FlaggedRevs was deployed, it retroactively put past edits made by the untrusted users under review, regardless of the date that the edit occurred. In this way, the system often required reviewers to assess a large volume of past contributions and choose whether to reject them. After the initial rollout, the number of reverted contributions among untrusted IP editors quickly drops even below the mean level, before gradually going back up. Statistical test result seen in Table ?? (see Appendix) of coefficients and confirm a significant impact of FlaggedRevs on the number of reverted contributions made by IP editors. We do not find the same effect on first-time registered editors (our second group of affected editors). While the number of rejected contributions shown in Figure 2 appears to increase gradually, this trend is consistent with the pre-intervention trend, suggesting that FlaggedRevs is unlikely to play a role in the uptrend. We also do not see any major effect of FlaggedRevs on the unaffected group of editors, as well as the overall community.

The impact of FlaggedRevs on the reverted rate of contributions shown in Figure 3 made by different group of users tells a slightly different story. We see neither a significant immediate change nor a change in trajectory caused by the intervention. While we do see a slight up uptick in the reverted rate during the month that FlaggedRevs is deployed, the rate quickly goes back down to the mean level before gradually increasing over time. Although nuanced in some ways, our overall results for H2 are clear and reflect a consistent null result in regards to our hypotheses: we find find little evidence of a major impact of the pre-moderation system on the quality of contributions.

Regarding our hypotheses H3, we again see two different stories for the two editor groups that are affected by FlaggedRevs. The deployment of the systems appears to be associated with an immediate decline in the number of contributions made by IP editors. Once again, we do not see a similar effect within the first-time registered editor group also affected by FlaggedRevs. While the system did not cause an immediate change in number of contributions among the returning registered editor group and the all editor group, it does appear associated with significant change in the trajectory compared to the pre-intervention period. That said, it does not correspond to negative growth (the post-intervention edit trend remains fairly flat and hovers near the mean level). The ITS regression results shown in Table 5 confirm that the visually apparent effects

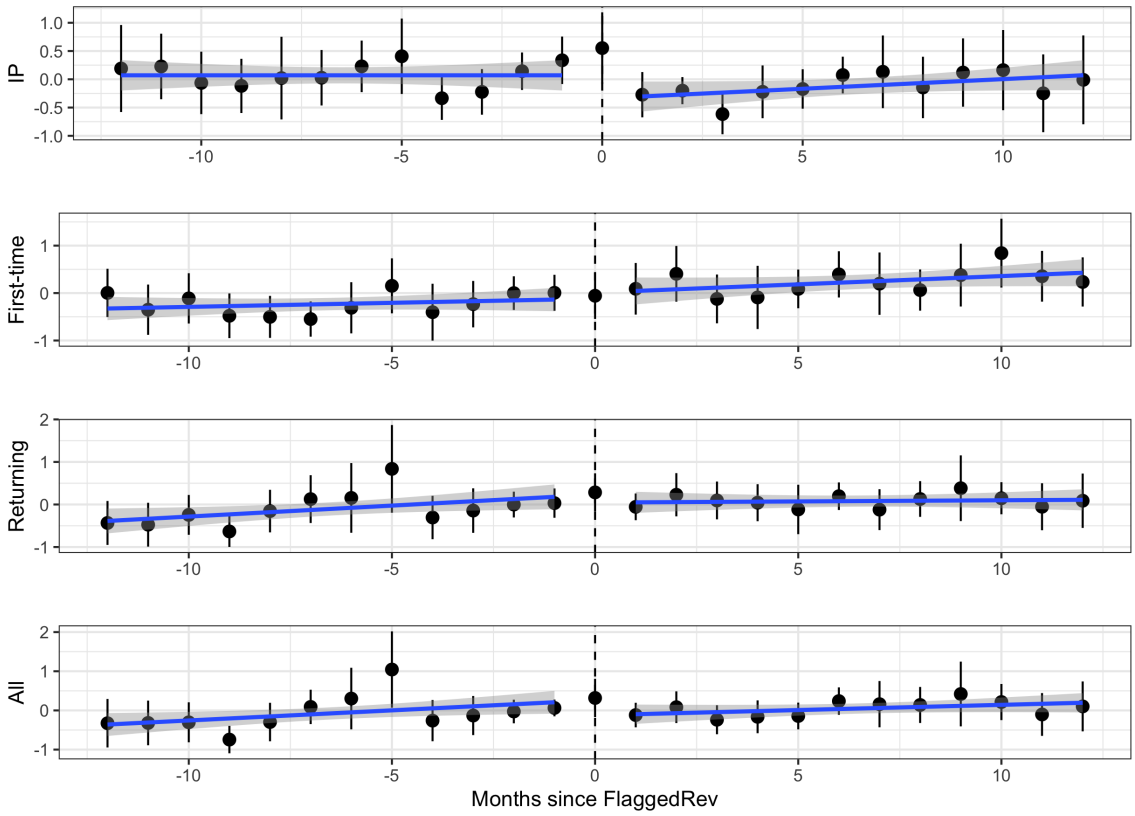

Fig. 2. Impact of FlaggedRevs on the number of reverted contributions (in std unit) made by different groups of users on Wikipedia.

described above are all statistically significant. Overall, we see the deployment of the pre-publication discouraged participation of the group of editors with lowest commitment and most targeted by the additional safeguard, but not other groups.

Finally, our test of H4 is reported in Table 1 which reports the estimates from our GLMM estimating newcomer return rate ~~on wikis that implement FlaggedRevs~~. We find that that although FlaggedRevs did negatively affected the return rate of newcomers in a way that was statistically significant, ~~its~~ effect ~~appears extremely~~ limited. Because

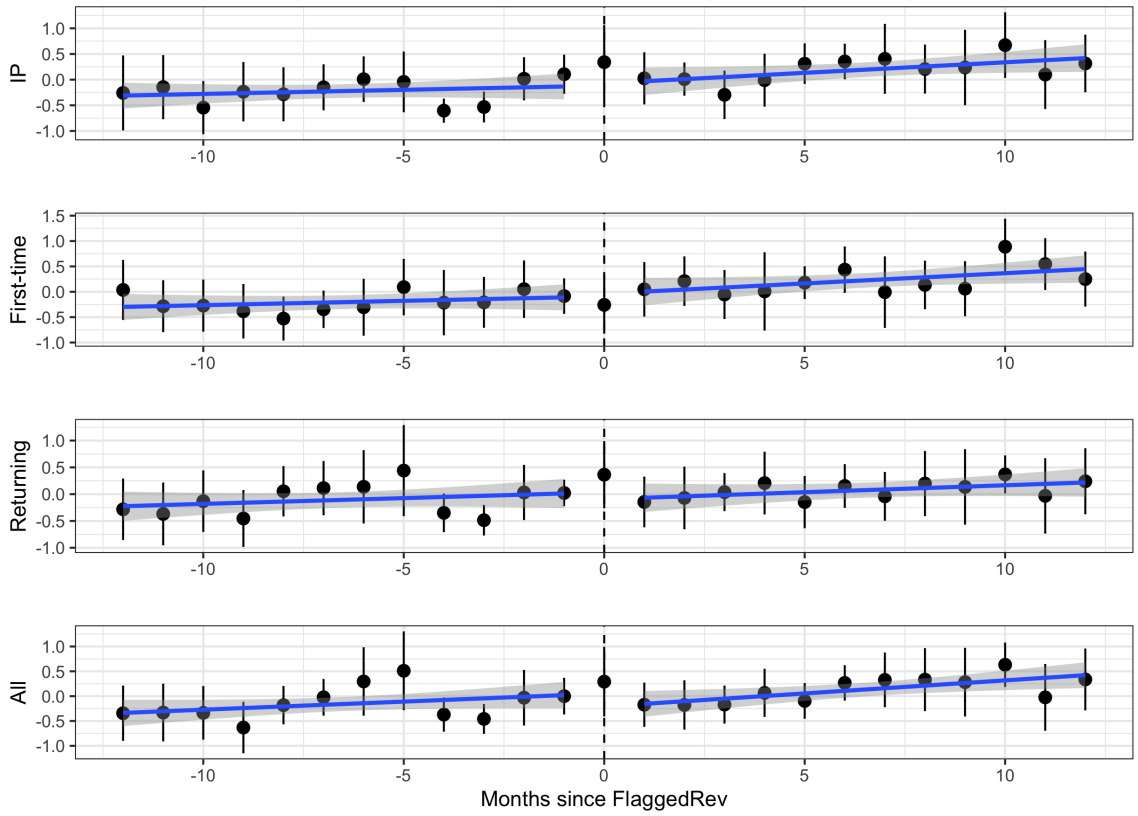

Fig. 3. Impact of FlaggedRevs on the reverted rate (in std unit) of contributions made by different groups of users on Wikipedia.

three way interactions can be difficult to interpret, we report a range of predicted values from our models and visualize these results in Figure 5. Our models suggests that before FlaggedRevs was enabled, 67% of newcomers return to make another edit session within 60 days of their first given that their first edit session was not published and that they registered for an account. Our model suggests that the return rate is only reduced by 2% on average ~~when~~ after FlaggedRevs is enabled.

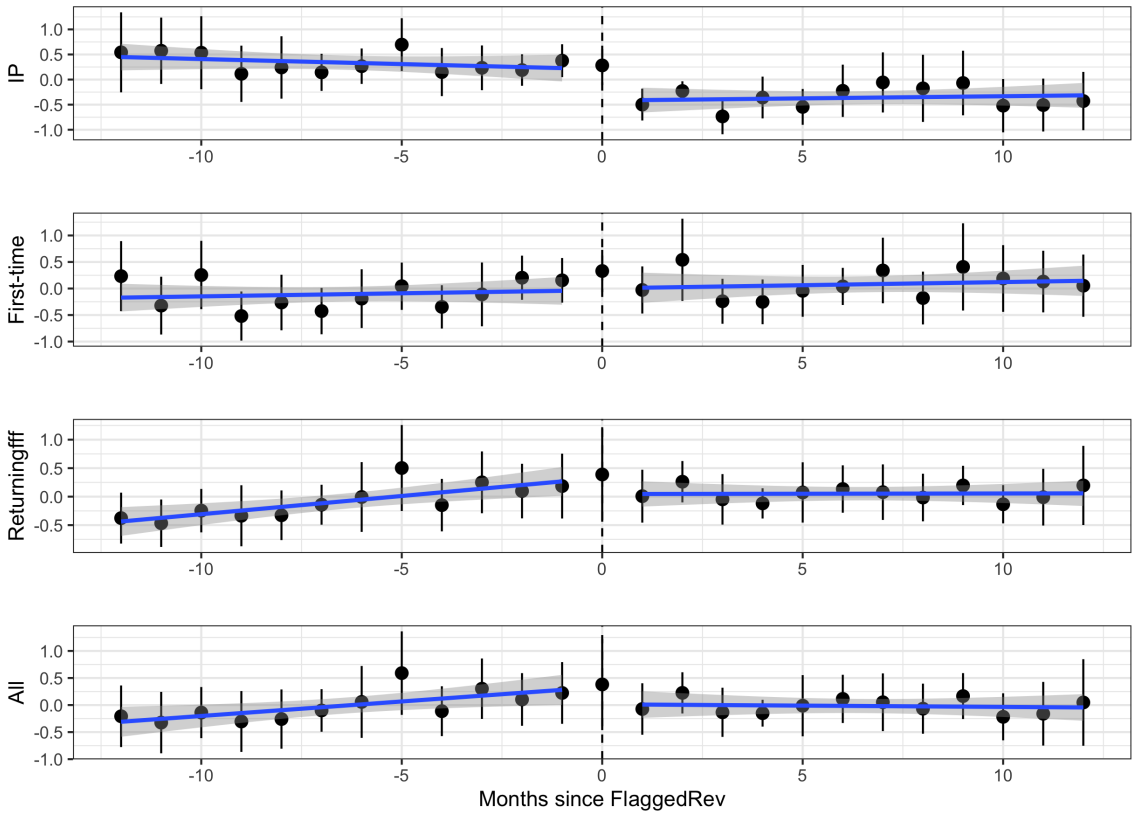

Fig. 4. Impact of FlaggedRevs on the number of contributions among different groups of users on Wikipedia.

This relatively small effect is a stark contrast to the effects of other variables in the same model. A newcomer's first edit session being published is associated with a large increase in the chance that the editor makes another edit session from 67% to 88%. We also find that first-time registered users are much more likely to return than first-time editors editing without accounts: first-time IP editors return at 15% rate when their first effort is not published and approximately 23% when it is published, regardless of whether or not

Table 1. Estimated values of the multilevel logistic model estimating maximum likelihood of newcomer return rate.

	Estimated Value	Std. Error	p-value
(Intercept)	0.771	0.159	<0.001 (***)
flaggedrevs_on	-0.044	0.020	0.031 (*)
first_edit_published	1.273	0.014	<0.001 (***)
ip	-2.47	0.016	<0.001 (***)
time	-0.015	0.006	0.024 (*)
first_edit_published& ip	-0.915	0.017	<0.001 (***)
flaggedrevs& first_edit_published	0.043	0.022	0.053 (.)
flaggedrevs& first_edit_published& ip	-0.056	0.024	0.020 (*)
Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' Number of observations: 1,972,861			

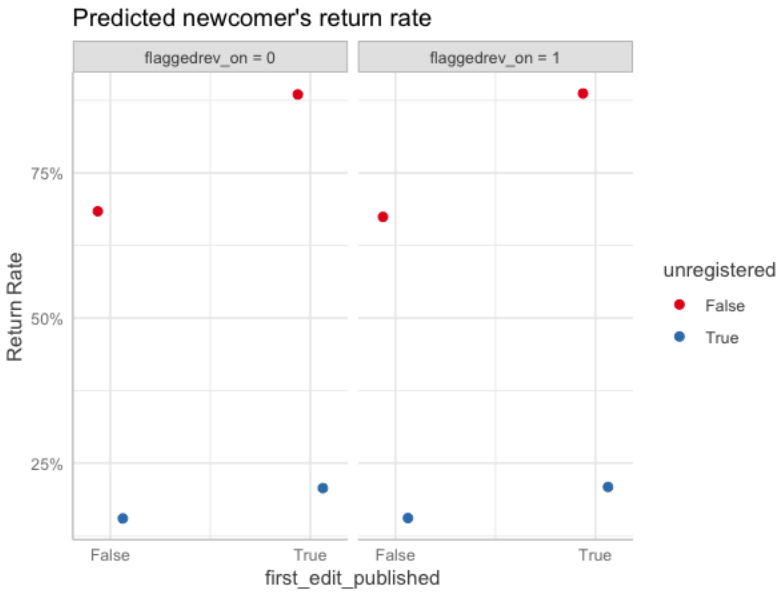

Fig. 5. Three-way interaction of different independent variables on newcomer return rate on wikis that implemented FlaggedRevs.

~~FlaggedRevs has been implemented. The return rate does not seem to vary significantly through time.~~

6 DISCUSSION AND LIMITATIONS

Our findings suggest that the deployment of pre-publication moderation did not result in major negative consequences on the quality, productivity, and sustainability of communities. These results will likely be seen as vindication for communities that have implemented FlaggedRevs and systems like it over objections that the collateral damage from the system would overwhelm its benefits. Our results provide little evidence to support these claims.

On the other hand, there a range of other reasons that Wikipedia language editions opt not to use pre-publication review that might still obtain. From a technical perspective, FlaggedRevs' source code lacks maintenance and development. Additionally, the system requires a set of configuration choices that must be customized to each community. All of these make it difficult for new wikis to easily adopt FlaggedRes and leads to frustrations among both members of communities that have deployed it. FlaggedRevs itself suffers from a range of specific limitations. For example, FlaggedRevs do not provide any notification for editors to let them know if their contribution has been rejected or approved.[13] Some users with review rights complained that "[u]ser interface is not made for checking all the edits in realtime".[14] Since April 2017, requests for deployment of the system by other wikis have been paused by the Wikimedia Foundation indefinitely for

[13] https://meta.wikimedia.org/wiki/Talk:Flagged_Revisions
[14] https://meta.wikimedia.org/wiki/Requests_for_comment/Flagged_revisions_deployment

these reasons.[15] In many senses, these challenges with the FlaggedRevs' system make its relative minimal effects on communities even more surprising.

An additional limitation that stems from the heterogeneity in communities deploying the system is that the experience of FlaggedRevs may vary enormously across communities. For example, wikis of vastly different sizes also have vastly different numbers of editors with review rights, leading to vastly different average review times. For example, German Wikipedia currently has 19,994 users with review rights, and the average waiting time of the pages with pending changes is 13 to 18 hours.[16] Meanwhile, Russian Wikipedia has 2,422 users with review rights, and the average waiting time of the pages with pending changes is 761 days 8 hours.[17]

At a minimum, we believe that our work suggests that, a streamlined version of FlaggedRevs could serve as a path for reviewing contributions from populations of editors that are currently deemed too "high risk" to contribute to peer production systems at all and have already been blocked across the board. For example, research by Chau et al. has shown that contributions from anonymity-seeking Tor users that are currently blocked from contributing to Wikipedia have been the source of substantial value in the past [49]. We believe that our study suggests that a pre-publication moderation system

[15] https://phabricator.wikimedia.org/T66726#3189794
[16] https://perma.cc/PL7G-M2ZE
[17] https://perma.cc/B3EV-82BC

like FlaggedRevs may be an effective way to ~~effectively~~ filter low quality contributions while retaining good-faith attempts.

7 CONCLUSION

This study seeks to measure the risks, benefits and unintended effects associated with the deployment of a pre-publication moderation system. The paper does so by presenting a case study of FlaggedRevs, a system ~~which~~ is designed by German Wikipedia and deployed by 24 Wikipedia language editions. First, we sought to understand if the deployment of FlaggedRevs caused a fundamental difference in the way contributions are reviewed and found that FlaggedRevs effectivedly prevented vandalism and other low quality contributions from ever being published. We did not find strong evidence of any meaningful long-term change in contribution quality.

We also found that, while the system caused a significant drop in the number of contributions made by IP editors, it did not impact users with accounts editing for the first time or users editing with accounts. ~~We did not estimate an overall community-level effect on contribution rates~~. Although we hypothesized that pre-publication moderation system would impact newcomers' return rate, we ~~once again~~ find that FlaggedRevs does not appear to have made a substantial impact (although the effect was negative and statistically significant). Users appeared to be largely unfazed by the delayed feedback.

Our results suggest that pre-publication review can both protect peer produced resources against bad contributions going public and without necessarily deterring participation.

APPENDIX

Table 2. Coefficients of the OLS model estimating the number of visible reverted contributions (in standard deviation unit) for each group of editors, associated with H1.

	Coefficient	Std. Error	p-value
IP Editors			
flaggedrev_on	-1.78	0.086	<0.001(***)
flaggedrev_on × *time*	0.014	0.011	0.238
First-time Editors			
flaggedrev_on	-1.759	0.095	<0.001(***)
flaggedrev_on × *time*	0.018	0.013	0.156
Returning Registered Editors			
flaggedrev_on	-0.348	0.188	0.065
flaggedrev_on × *time*	-0.056	0.026	0.129
All Editors			
flaggedrev_on	-1.27	0.167	<0.001(***)
flaggedrev_on × *time*	-0.035	0.023	0.124

Table 3. Coefficients of the OLS model estimating the number of reverted contributions (in standard deviation unit) for each group of editors, associated with H2.

	Coefficient	Std. Error	p-value
IP Editors			
flaggedrev_on	-0.54	0.203	0.008(**)
flaggedrev_on × *time*	0.018	0.028	0.517
First-time Editors			
flaggedrev_on	0.019	0.213	0.640
flaggedrev_on × *time*	0.015	0.027	0.57
Returning Registered Editors			
flaggedrev_on	-0.20	0.202	0.322
flaggedrev_on × *time*	-0.047	0.028	0.088 (*)
All Editors			
flaggedrev_on	-0.399	0.202	0.056
flaggedrev_on × *time*	-0.026	0.028	0.336

Table 4. Coefficients of the OLS model estimating the reverted rate (in standard deviation unit) for each group of editors, associated with H2.

	Coefficient	Std. Error	p-value
IP Editors			
flaggedrev_on	-0.080	0.20	0.830
flaggedrev_on × *time*	0.009	0.027	0.723
First-time Editors			
flaggedrev_on	0.101	0.199	0.612
flaggedrev_on × *time*	0.028	0.027	0.299
Returning Registered Editors			
flaggedrev_on	-0.217	0.203	0.286
flaggedrev_on × *time*	-0.006	0.028	0.815
All Editors			
flaggedrev_on	-0.326	0.200	0.104
flaggedrev_on × *time*	0.012	0.027	0.664

Table 5. Coefficients of the OLS model estimating the number of edits (in standard deviation unit) made by each group of editors, associated with H3.

	Coefficient	Std. Error	p-value
IP Editors			
flaggedrev_on	-0.626	0.209	0.003(**)
flaggedrev_on × *time*	0.0287	0.028	0.3120
First-time Editors			
flaggedrev_on	0.0002	0.030	0.994
flaggedrev_on × *time*	0.0348	0.222	0.876
Returning Registered Editors			
flaggedrev_on	-0.285	0.192	0.1392
flaggedrev_on × *time*	-0.063	0.026	0.067 (*)
All Editors			
flaggedrev_on	-0.320	0.209	0.20
flaggedrev_on × *time*	-0.064	0.028	0.062 (*)

REFERENCES

[1] Denise Anthony, Sean W Smith, and Timothy Williamson. 2009. Reputation and reliability in collective goods: The case of the online encyclopedia Wikipedia. *Rationality and Society* 21, 3 (2009), 283–306.

[2] Judd Antin and Coye Cheshire. 2010. Readers Are Not Free-Riders: Reading as a Form of Participation on Wikipedia. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 127–130. https://doi.org/10.1145/1718918.1718942

[3] Yochai Benkler. 2002. Coase's penguin, or, Linux and 'The nature of the firm'. *The Yale Law Journal* 112, 3 (Dec. 2002), 369. https://doi.org/10.2307/1562247

[4] Yochai Benkler. 2006. *The wealth of networks: How social production transforms markets and freedom.* Yale University Press.

[5] Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2015. Peer production: A form of collective intelligence. In *Handbook of Collective Intelligence*, Thomas W. Malone and Michael S. Bernstein (Eds.). MIT Press, Cambridge, MA, 175–204.

[6] W. Lance Bennett and Alexandra Segerberg. 2012. The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication & Society* 15, 5 (2012), 739–768. https://doi.org/10.1080/1369118X.2012.670661

[7] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. 2016. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology* 46, 1 (06 2016), 348–355. https://doi.org/10.1093/ije/dyw098

[8] B Butler. 1999. *When is a group not a group: An empirical examination of metaphors for online social structure (chapter 1). The dynamics of cyberspace: Examining and modeling online social structure (pp 1-46).* Ph.D. Dissertation. Carnegie Mellon University, Pittsburgh, PA.

[9] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666

[10] Coye Cheshire. 2007. Selective Incentives and Generalized Information Exchange. *Social Psychology Quarterly* 70, 1 (2007), 82–100. http://www.jstor.org/stable/20141769

[11] Coye Cheshire and Judd Antin. 2008. The Social Psychological Effects of Feedback on the Production of Internet Information Pools. *Journal of Computer-Mediated Communication* 13, 3 (04 2008), 705–727. https://doi.org/10.1111/j.1083-6101.2008.00416.x

[12] Angela Gracia B Cruz, Yuri Seo, and Mathew Rex. 2018. Trolling in online communities: A practice-based theoretical perspective. *The Information Society* 34, 1 (2018), 15–26.

[13] R Geiger, Aaron Halfaker, Maryana Pinchuk, and Steven Walling. 2012. Defense mechanism or socialization tactic? Improving Wikipedia's notifications to rejected contributors. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6.

[14] R. Stuart Geiger and Aaron Halfaker. 2013. Using Edit Sessions to Measure Participation in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 861–870. https://doi.org/10.1145/2441776.2441873

[15] R Stuart Geiger and Aaron Halfaker. 2013. When the levee breaks: without bots, what happens to Wikipedia's quality control processes?. In *Proceedings of the 9th International Symposium on Open Collaboration*. 1–6.

[16] R. Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, 117–126. https://doi.org/10.1145/1718918.1718941

[17] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press, New Haven, Connecticut.

[18] Miriam Greis, Florian Alt, Niels Henze, and Nemanja Memarovic. 2014. I can wait a minute: uncovering the optimal delay time for pre-moderated user-generated content on public displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1435–1438.

[19] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–37. https://doi.org/10.1145/3415219

[20] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 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 (2013), 664–688. https://doi.org/10.1177/0002764212469365

[21] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. ACM, New York, NY, USA, 163–172. https://doi.org/10.1145/2038558.2038585

[22] Benjamin Mako Hill and Aaron Shaw. 2020. The Hidden Costs of Requiring Accounts: Quasi-Experimental Evidence From Peer Production. *Communication Research* (May 2020). https://doi.org/10.1177/0093650220910345

[23] Gerald C. Kane, Jeremiah Johnson, and Ann Majchrzak. 2014. Emergent Life Cycle: The Tension Between Knowledge Change and Knowledge Retention in Open Online Coproduction Communities. *Management Science* 60, 12 (2014).

[24] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 453–462. https://doi.org/10.1145/1240624.1240698

[25] Evangelos Kontopantelis, Tim Doran, David A Springate, Iain Buchan, and David Reeves. 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ* 350 (2015), h2750.

[26] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design.* MIT Press.

[27] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Montréal Québec Canada, 591–602. https://doi.org/10.1145/2872427.2883085

[28] Cliff Lampe and Erik Johnston. 2005. Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work (GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/1099203.1099206

[29] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 543–550.

[30] Martin Lea, T O'Shea, P Fung, and Russell Spears. 1992. Flaming" in computermediated-communication, volume Contexts of computer-mediated communication. *Harvester Wheatsheaf, Hertfordshire, England* 1 (1992), 89–112.

[31] Alex Leavitt. 2015. "This is a Throwaway Account" Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 317–327.

[32] Gerald Marwell and Pamela Oliver. 1993. *The critical mass in collective action.* Cambridge University Press.

[33] Charles E McCulloch and John M Neuhaus. 2014. Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online* (2014).

[34] Blair Nonnecke and Jenny Preece. 2001. Why lurkers lurk. *AMCIS 2001 proceedings* (2001), 294.

[35] Anthony Oberschall. 1973. *Social conflict and social movements.* Prentice-Hall Englewood Cliffs, NJ.

[36] Mancur Olson. 1989. Collective action. In *The Invisible Hand*. Springer, 61–69.

[37] Mancur Olson. 2012. The logic of collective action [1965]. *Contemporary Sociological Theory* 124 (2012).

[38] Felipe Ortega, Jesus M Gonzalez-Barahona, and Gregorio Robles. 2008. On the inequality of contributions to Wikipedia. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, 304–304.

[39] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 137 (Nov. 2018), 23 pages. https://doi.org/10.1145/3274406

[40] Robert B Penfold and Fang Zhang. 2013. Use of interrupted time series analysis in evaluating health care quality improvements. *Academic Pediatrics* 13, 6 (2013), S38–S44.

[41] Nathaniel Poor. 2005. Mechanisms of an online public sphere: The website Slashdot. *Journal of computer-mediated communication* 10, 2 (2005), JCMC1028.

[42] Jenny Preece. 2000. *Online communities: Designing usability, supporting sociability*. John Wiley, New York.

[43] Jodi Schneider, Bluma S. Gelley, and Aaron Halfaker. 2014. Accept, Decline, Postpone: How Newcomer Productivity is Reduced in English Wikipedia by Pre-Publication Review. In *Proceedings of The International Symposium on Open Collaboration (OpenSym '14)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2641580.2641614

[44] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[45] Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36, 3 (2010), 357–370. https://doi.org/10.1177/0165551510365390

[46] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation Strategy: Compliance and other outcomes in the ChangeMyView community. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 163 (Nov. 2019), 21 pages. https://doi.org/10.1145/3359265

[47] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. The Singularity is Not near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)*. Association for Computing Machinery, New York, NY, USA, Article 8, 10 pages. https://doi.org/10.1145/1641309.1641322

[48] Nathan TeBlunthuis, Aaron Shaw, and Benjamin Mako Hill. 2018. Revisiting "The Rise and Decline" in a Population of Peer Production Projects. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173929

[49] C. Tran, K. Champion, A. Forte, B. M. Hill, and R. Greenstadt. 2020. Are anonymity-seekers just like everybody else? An analysis of contributions to Wikipedia from Tor. In *2020 IEEE Symposium on Security and Privacy (SP)*. 186–202.

[50] Karin Wahl-Jorgensen, Andrew Williams, and Claire Wardle. 2010. Audience views on user-generated content: Exploring the value of news from the bottom up. *Northern Lights: Film & Media Studies Yearbook* 8, 1 (2010), 177–194.

[51] Kevin Wise, Brian Hamman, and Kjerstin Thorson. 2006. Moderation, Response Rate, and Message Interactivity: Features of Online Communities and Their Effects on Intent to Participate. *Journal of Computer-Mediated Communication* 12, 1 (10 2006), 24–41. https://doi.org/10.1111/j.1083-6101.2006.00313.x