Etiquette in Wikipedia: Weening New Editors into Productive Ones

Ryan Faulkner
Wikimedia Foundation
149 New Montgomery St.
San Francisco, California USA
rfaulkner@wikimedia.org

Steven Walling
Wikimedia Foundation
149 New Montgomery St.
San Francisco, California USA
swalling@wikimedia.org

Maryana Pinchuk
Wikimedia Foundation
149 New Montgomery St.
San Francisco, California USA
mpinchuk@wikimedia.org

ABSTRACT

Currently, the greatest challenge faced by the Wikipedia community involves reversing the decline of active editors on the site – in other words, ensuring that the encyclopedia's contributors remain sufficiently numerous to fill the roles that keep it relevant. Due to the natural drop-off of old contributors, newcomers must constantly be socialized, trained and retained. However recent research has shown the Wikipedia community is failing to retain a large proportion of productive new contributors and implicates Wikipedia's semi-automated quality control mechanisms and their interactions with these newcomers as an exacerbating factor. This paper evaluates the effectiveness of minor changes to the normative warning messages sent to newcomers from one of the most prolific of these quality control tools (Huggle) in preserving their rate of contribution. The experimental results suggest that substantial gains in newcomer participation can be attained through inexpensive changes to the wording of the first normative message that new contributors receive.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Wikipedia, Huggle, Newcomers, Vandalism, Wiki, Retention

1. INTRODUCTION

For open collaboration communities like Wikipedia, maintaining a healthy community of volunteer contributors is essential to remaining relevant. Historically, English Wikipedia has been highly successful in this regard, but recent research has shown that the growth in the online encyclopedia's community of editors suddenly halted in early 2007 and has since entered a steady decline [6]. This research identified a decrease in the rate of newcomer retention as the root cause, and the "Editor Trends Study" [1], conducted by the Wikimedia Foundation, supports this conclusion.

A shrinking community of editors could negatively impact the health of Wikipedia in several ways, most notably by the reduction in the diversity and volume of quality content and on the maintenance of the project as a whole. If the trend continues, the project may fall below some critical threshold of active contributors; beneath which, the encyclopedia might rapidly decrease in quality and relevance. This makes retention of new editors a particularly interesting and vital problem to the hundreds of millions of consumers of Wikipedia's encyclopedia, the community of editors, and the Wikimedia Foundation.

1.1 New editor experience

Wikipedia's primary quality control mechanism relies on the ability of editors to reject damaging contributions by reverting articles to a state before the damage was made. Early in the career of an editor, there is a good chance that at least some portion of their edits will be reverted by more experienced editors [5]. Although this is often due to the intentionally damaging nature of the contributions (as might be expected for an open encyclopedia), many reverts come in response to contributions that violate Wikipedia norms and contribution guidelines that newcomers are unfamiliar with. However, differentiating between "bad-faith" vandalism and a "good-faith" mistake is a not always easy, especially early on in an editor's career.

Along with the revert of damaging content, experienced editors will typically send a warning message to the violator by appending it on the editor's "Talk" page. Messages of this form are often delivered using Wikipedia templates[1], a system that allows editors to quickly create a standardized message with information personalized to the location and nature of the originating revert.

1.2 Revert tools & warnings

In order to deal with the massive rate of article edits that occur in Wikipedia, editors have constructed tools that help detect contributions made in violation of Wikipedia policy, and that may also subsequently warn editors about these infractions in an automated or semi-automated manner. These tools come in two general forms: (1) "bots"[2], fully automated computer programs that crawl Wikipedia to identify and automatically revert the most obviously damaging edits; (2) "Power Tools", application software with the

WikiSym'12 Linz, Austria

[1] `http://en.wikipedia.org/wiki/Help:Template`
[2] `http://en.wikipedia.org/wiki/Wikipedia:Bots`

express purpose of aiding the user of the tool in finding instances of vandalism, whereupon they may revert an edit and subsequently warn the author. Currently, there are two prominent power tools for performing reverts in Wikipedia: “Twinkle” and “Huggle”[3]. Unlike bots, these tools use human judgement and are thus able to identify a much wider range of edits. In this paper, we’ll focus on Huggle, since it is the most prolific of the two tools.

As suggested above, when a user of Huggle reverts an edit, it will also append a template-based warning to the reverted editor’s Talk page. In the case that an editor is reverted more than once in quick succession Huggle will leave increasingly serious warnings, where the strength of the tone of the reprimand is increased with each additional warning, finally culminating in a direct warning to the editor, that if they continue to vandalize they risk being blocked from Wikipedia. These are labeled warning “levels”, where the first warning received is defined as a level-1 warning, the second as a level-2 warning, and so on (see [3] for an overview). However, of most interest to an exploration of retaining good editors are the level-1 warnings, since these messages are delivered when the editor is at their most novice and is therefore more likely to be simply unaware of policy violations on which their contributions may infringe.

Wikipedia guidelines encourage editors to “Assume Good Faith”[4]. In other words, when there is doubt, assume that the editor is at least trying to be helpful. The first warning is a prime opportunity for the application of this guideline.

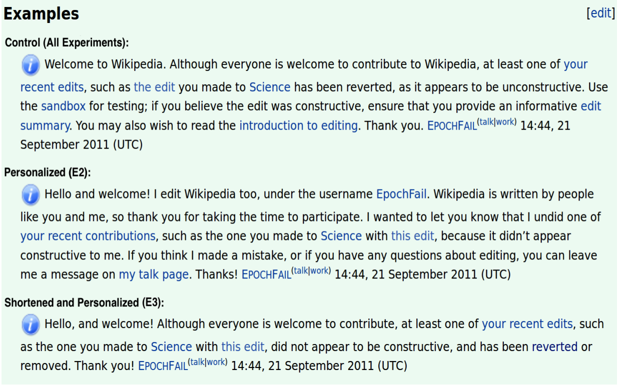
1.3 Huggle’s newcomer warnings

In English Wikipedia, there are on average roughly 1,000 new user accounts registered per day that save at least one edit [7] and approximately 20,000 first warning messages delivered to new accounts per month [2]. Of these, approximately 80% of all first warning messages are delivered by bots or power tools, where Huggle has been responsible for the delivery of 10-40% of total first warnings in any given month from 2008 to mid-2011 [2]. However it should be noted, that the method of delivery of a warning template for vandalism is not of particular concern as our focus involves the effect of the message itself on a new editor’s experience, and therefore what may be learned about first warnings in Huggle should be applicable to other power tools.

Clearly, semi-automated warnings are ubiquitous in English Wikipedia, but they are also a potentially significant mechanism for editor growth. Huggle is the only semi-automated tool that deals solely with policy infraction (this is a superset of vandalism) by editors, while also accounting for a significant portion of first warnings, therefore Huggle is a sensible choice for experimentation on first warning messages to new editors. The focus of the experiments will be to determine if the retention and productivity of new editors can be affected by modifying the first warning message that new editors receive. Our hypothesis is the following:

> *Friendlier and clearer templates generated by vandal fighting tools can increase the productivity of new editors.*

The following section will detail the experiments and analysis that was carried out to test this hypothesis. It is important to mention that an earlier iteration of this work can be found in [4], the experiments in this paper are a direct extension in several ways. First a new type of message is tested, namely messages that were reduced in length, or “shortened” messages. Second, issue specific warnings were tested and finally, a new metric: *normalized edit difference*, is measured. This metric includes information about editor contributions both before and after a warning is received. These extensions are defined in the next section.



Examples [edit]

Control (All Experiments):

Welcome to Wikipedia. Although everyone is welcome to contribute to Wikipedia, at least one of your recent edits, such as the edit you made to Science has been reverted, as it appears to be unconstructive. Use the sandbox for testing; if you believe the edit was constructive, ensure that you provide an informative edit summary. You may also wish to read the introduction to editing. Thank you. EPOCHFAIL(talk|work) 14:44, 21 September 2011 (UTC)

Personalized (E2):

Hello and welcome! I edit Wikipedia too, under the username EpochFail. Wikipedia is written by people like you and me, so thank you for taking the time to participate. I wanted to let you know that I undid one of your recent contributions, such as the one you made to Science with this edit, because it didn’t appear constructive to me. If you think I made a mistake, or if you have any questions about editing, you can leave me a message on my talk page. Thanks! EPOCHFAIL(talk|work) 14:44, 21 September 2011 (UTC)

Shortened and Personalized (E3):

Hello, and welcome! Although everyone is welcome to contribute, at least one of your recent edits, such as the one you made to Science with this edit, did not appear to be constructive, and has been reverted or removed. Thank you! EPOCHFAIL(talk|work) 14:44, 21 September 2011 (UTC)

Figure 1: An example of a “shortened” message, a “personalized” message, and a standard warning template message delivered by Huggle.

2. EXPERIMENTS & METHODOLOGY

To determine the validity of the hypothesis, it was necessary to create a neutral control message and one or more test messages communicating the desired effect. Examples of these template flavors can be seen in Figure 1. Templates were altered in two different ways, 1) by “personalizing” them (making the tone of the message more informal, less accusatory and more passive, and removing explicit directives linking to policy pages), and 2) by “shortening” the warning message, making the language of the warning simpler and more understandable to the recipient.

To measure how new editors react to different warnings it was necessary to configure Huggle to deliver templates based on a switch parser function[5] that chooses templates at random. This modification did not result in any functional change for editors using the tool. Templates corresponding to warnings were chosen at random among existing warning templates (this includes non-experimental templates) with no knowledge of which version was being sent to the target editor. This blind delivery ensured that the editor groups that received the test and control templates did not suffer from selection bias and were of approximately the same size. Editors receving the templates met the following criteria: (1) the warning is a first warning (or a level-1 warning), (2) the editor must not go on to be blocked after receiving the warning, (3) the editor would need to be a registered user on Wikipedia (i.e. not anonymous). This ensured that the revision (“revision” can be taken as equivalent to the notion of an “edit”) for which the editor was warned was likely to be an edit made in good faith. These experiments ran in Huggle from Nov. 8th, 2011 to Dec. 9th, 2011 - the time periods

[3] `http://en.wikipedia.org/wiki/Wikipedia:Huggle`

[4] `http://en.wikipedia.org/wiki/Wikipedia:Assume_good_faith`

[5] `https://meta.wikimedia.org/wiki/Parser_function`

varied by experiment and did not completely overlap.

In addition to the generic vandalism warnings, Huggle users have the option to also specify a particular type of warning to issue to an editor, dependent upon the violation associated with the revision. Experimental messages were created for the following specific policy violations:

- *test edits* - addition of an edit as a test (not content)
- *spamming* - addition of an external link to the body of an article
- *unsourced content* - new content added to an article without a clear source
- *deletion* - removal of a portion of the content of an article without explanation or a clear reason

To monitor these experiements, a clone of the production database powering English Wikipedia provided access to editor data, revision comments and activity, editor warning events, and editor block events. The MediaWiki API[6] was used to find specific portions of text that corresponded to the experimental templates delivered via Huggle, allowing for positive identification of the sample data. With these sources the following measurements were compiled for each editor that received an experimental warning from Huggle: (1) the exact time each experimental template was issued, (2) the editor that received the template on his or her Talk page, (3) the number of revisions in all namespaces over the editor's lifetime before the warning template was issued and in the three day period after, (4) the number of warnings in the editor's lifetime before the warning template was issued and in the three day period after, and (5) the number of blocks in the editor's lifetime before the warning template was issued and in the three day period after.

Measuring revision activity over a three day period following the warning was chosen empirically after observing the window that often yielded the largest effect on editing behavior. This period provided a balance that allowed a significant number of revisions to accumulate on average for most editors, while still allowing for the observation of a significant effect from the warning.

Editors were further bucketed into overlapping groups based on the **minimum** number of edits they made before receiving a warning. Therefore, an editor is said to belong to an *editor group* G_n if they have made **at least** n edits before receiving a warning from an experimental template. For each editor u, let $j_{before}(u)$ be the number of revisions in all namespaces before the warning, and let $j_{after}(u)$ be the number of revisions in all namespaces in the three day period following the receipt of the warning. From these values the *normalized edit difference* can be measured, that is the difference between $j_{before}(u)$ and $j_{after}(u)$ normalized by $j_{before}(u)$ (see $m(u)$ in the formula below). This seemed a suitable metric as it provides a measure of the reduction in the edit activity of an editor after being warned (hence lower values are better), but also normalizes this value with respect to the amount of activity the editor engaged in before the warning. Once these quantities were measured for each editor, logistic regression was used to measure the response among the warning message groups for *normalized edit difference* – the regression model is presented in the

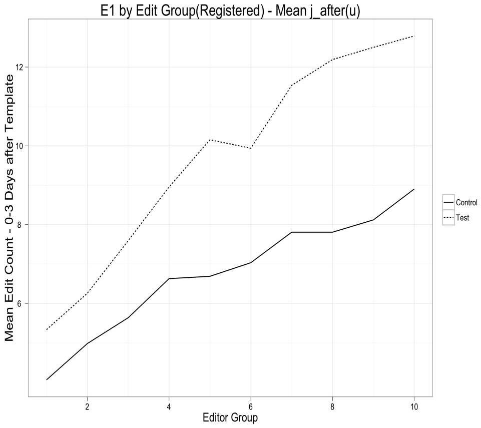


Figure 2: $j_{after}(u)$ after receiving the "shortened" warning (E1) as a function of editor groups.

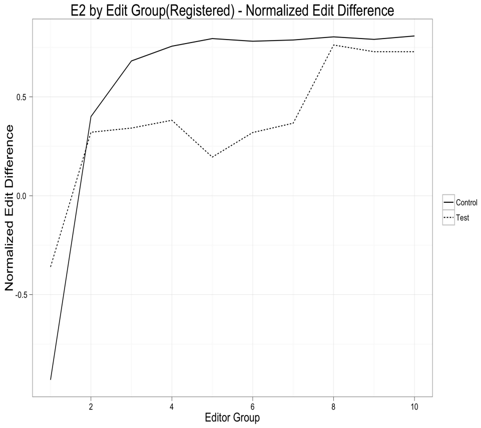


Figure 3: Normalized edit difference (E2) as a function of editor groups.

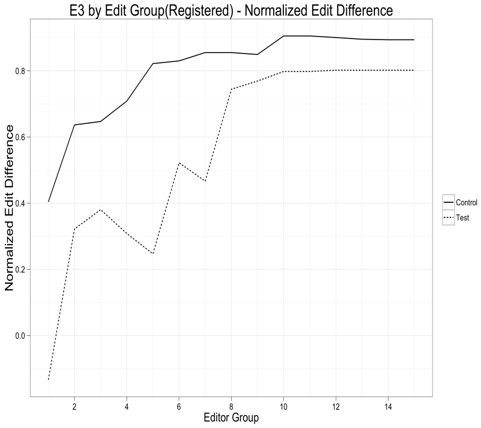


Figure 4: Normalized edit difference (E3) as a function of editor groups.

[6] `http://www.mediawiki.org/wiki/API:Main_page`

Table 1: Results of logistic regression on *normalized edit difference*

Experiment	test sample	control sample	Editor Group	β_1	error	p-value	AIC
E1: shortened warning	26	44	G_5	-2.151	0.9864	0.0315	88.418
E2: personalized warning	29	35	G_5	-1.496	0.5577	0.135	85.815
E3: shortened & personalized warning	32	32	G_5	-1.4906	0.5964	0.0124	89.088

formula (below), $template(u)$ is a binary variable assigned a value based on the template seen by the editor.

$$m(u) = (j_{before}(u) - j_{after}(u))/j_{before}(u)$$

$$logit(template(u)) = \beta_0 + \beta_1 * m(u).$$

The experiments are listed in table 1 alongside the results of the regression analysis. The first experiment, **E1**, involves shortened messages and applies to all edits which violate policy but are not classified as deliberate infractions (vandalism). The shortened message is a reduced length message in a fashion similar to that seen in Figure 1. The second experiment, **E2**, tests the effect of a "personalized" message (an example can also be seen in Figure 1), that is, a message which is sympathetic with the target editor, explicitly acknowledging that the edit was one made in good faith. The final experiment, **E3**, combines both a personalized and shortened warning, and also tests issue specific warnings[7].

3. RESULTS & DISCUSSION

Figures 3 and 4 show a positive effect on the *normalized edit difference* of new editors for both of the modified warning messages. There appears to be a "sweet spot" between two and ten edits where the difference in performance between editors who received the control and experimental message was the largest. This result could suggest that new editors are more receptive to these types of messages. In Figures 3 and 4 in particular we can see this area among the editor groups (G_2 through G_{10}) where the test outperforms the control more definitively under the mean *normalized edit difference*. This is not entirely surprising since warnings may tend to exert less of an influence on an editor as they become more experienced. However, the shortened templates did do very well over a larger range of editor groups as can be seen in Figure 2 for $j_{after}(u)$; this was observed for *normalized edit difference* also. The analysis in the previous section was also repeated over $j_{after}(u)$, instead of *normalized edit difference*, with similar results. Figure 2 shows the comparison of the experimental groups as a function of editor groups for experiment **E1** from Table 1. Higher levels of edit activity, after receiving the warning, was on average predictive of an editor having received the modified warning message.

Table 1 shows experiments yielded results ranging from marginally-significant to significant outcomes in favor of shortening and personalizing first time warning messages to new editors having at least four or five prior edits. Therefore, by modifying first time warnings, the likelihood that a novice editor contributes in productive ways, in the immediate timeframe following the posting, can be significantly increased. The analysis of this paper implies that there is clear merit to leveraging semi-automated approaches to vandal fighting in two ways: (1) to stimulate the productivity of new editors by framing their early interaction with the community in the context of concise and friendly feedback, and (2) to facilitate experimental work with the community that can lead us to learn more about them. Additionally, since this work was carried out on such a widely used resource as English Wikipedia, these results might be applied to collaborative systems in general with similar communities and feedback mechanisms. As a follow up to this work, there is both a push to engage the community in an effort to begin making first time warnings to editors more direct and personalized. Future work may also involve the investigation of a wider set of treatments among test templates and a measurement of longer term effects of these messages on editors, specifically, that if new editors are stimulated to be more productive in the short term, then this may lead to an increased chance of becoming a long term productive editor, or "Wikipedian".

4. ACKNOWLEDGMENTS

Thanks to A. Halfaker, S. Geiger, D. Taraborelli, and the Huggle development team for their assistance in providing the tools and review that informed this research.

5. REFERENCES

[1] W. Foundation. Editor trends study. http://strategy.wikimedia.org/wiki/Editor_Trends_Study, October 2010.

[2] W. Foundation. Summer of research summary of findings. https://meta.wikimedia.org/wiki/Research:Wikimedia_Summer_of_Research_2011/Summary_of_Findings, September 2011.

[3] R. S. Geiger and D. Ribes. 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, pages 117–126, New York, NY, USA, 2010. ACM.

[4] S. Geiger, A. Halfaker, M. Pinchuk, and S. Walling. Defense mechanism or socialization tactic? improving wikipedia's notifications to rejected contributors. In *The 6th International AAAI Conference ON Weblogs And Social Media*, ICWSM '12, 2012.

[5] A. Halfaker, A. Kittur, and J. Riedl. 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, pages 163–172, New York, NY, USA, 2011. ACM.

[6] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: slowing growth of wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 8:1–8:10, New York, NY, USA, 2009. ACM.

[7] D. Taraborelli. New user registrations. http://toolserver.org/~dartar/reg2/, July 2012.

[7] This experiment only involved editors receiving warning types: *test*, *delete*, *spam*, *unsourced*.