The Success and Failure of Quality Improvement Projects in Peer Production Communities

Morten Warncke-Wang, Vladislav R. Ayukaev, Brent Hecht, and Loren Terveen

GroupLens Research
University of Minnesota
Minneapolis, MN 55455, USA
{morten,ayukaev,bhecht,terveen}@cs.umn.edu

ABSTRACT

Peer production communities have been proven to be successful at creating valuable artefacts, with Wikipedia as a prime example. However, a number of studies have shown that work in these communities tends to be of uneven quality and certain content areas receive more attention than others. In this paper, we examine the efficacy of a range of targeted strategies to increase the quality of under-attended content areas in peer production communities. Mining data from five quality improvement projects in the English Wikipedia, the largest peer production community in the world, we show that certain types of strategies (e.g. creating artefacts from scratch) have better quality outcomes than others (e.g. improving existing artefacts), even if both are done by a similar cohort of participants. We discuss the implications of our findings for Wikipedia as well as other peer production communities.

Author Keywords

peer production; user-generated content; quality modelling; Wikipedia

ACM Classification Keywords

H.5.3. Information Interfaces: Group and Organization Interfaces – Computer-supported cooperative work

INTRODUCTION

Peer production communities like Wikipedia and OpenStreetMap have been successful at generating large quantities of artefacts, but coverage and content quality of certain subjects can be poor. Popular media reported on Wikipedia's gender bias[1] and argued that it leads to less content about topics with a female audience, a hypothesis that Lam et al. confirmed for articles about movies [23]. Research on biographies comparing Wikipedia to Encyclopædia Britannica found that the former had better coverage and longer articles but was also more likely to be missing articles about women [33], and Wikipedia's related problems with categorisation of novelists has attracted media attention[2]. This type of problem is not limited to Wikipedia: studies of contributions to OpenStreetMap found a much larger variety of tags available to describe places related to prostitution than those related to child care [37], and lower-income areas have considerably lower coverage and content quality [14].

Solutions to these problems have been proposed both from within the communities themselves, as in groups of Wikipedia editors self-organising to improve content (called "WikiProjects") about women scientists[3] and artists[4], and also through formally organised efforts. The Wikimedia Foundation manages the Wikipedia Education Program[5], a project where educators and students across the globe work on improving Wikipedia articles as class assignments.

But the question is: do these "solutions" actually work? Some of these projects have been studied in isolation, for instance the Association for Psychological Science's (APS) Wikipedia Initiative[6], a part of the Education Program, was studied by Farzan and Kraut [10], and WikiProjects were studied by Zhu, Kittur, and Kraut [45]. However, each study took a different approach and used different measurements: Zhu et al. measured amount of effort by counting number of edits made to articles, while Farzan and Kraut measured quality by counting words added and quantifying word survival. Similarly, the Wikimedia Foundation measured quality of work done in the Education Program using human assessments of random selections of articles [34, 43].

Our research begins the process of developing a coherent framework to describe, analyse, and evaluate quality improvement projects for peer production communities. We study five different projects in the English Wikipedia, the largest peer production community in the world, to identify the factors and mechanisms associated with successful projects, resulting in the following three major findings:

[1] `http://www.nytimes.com/2011/01/31/business/media/31link.html?_r=0`

[2] `http://www.nytimes.com/2013/04/28/opinion/sunday/wikipedias-sexism-toward-female-novelists.html`

[3] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_scientists`

[4] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_artists`

[5] `http://outreach.wikimedia.org/wiki/Education`

[6] `http://www.psychologicalscience.org/index.php/members/aps-wikipedia-initiative`

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CSCW '15, March 14–18, 2015, Vancouver, BC, Canada
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2922-4/15/03... $15.00
http://dx.doi.org/10.1145/2675133.2675241

1. Projects where participants work individually or in small groups tend to be more effective; an increased number of participants working on each artefact is connected with lower rates of quality increase.

2. Creating new artefacts appears to result in higher quality content than improving existing artefacts.

3. Simply drawing attention to a community's need for quality improvement is not associated with quality improvements; incentives or task structuring also are likely needed.

This work binds several threads of research together to advance how we describe and measure quality improvement projects, as well as our understanding of what factors lead to project success. Our results have implications for both volunteer and professional peer production communities. We begin by looking at related research to show how this paper builds on and extends the current state of knowledge.

RELATED WORK

The motivation for this work arises from two areas of peer production community research: increasing contributions, and quality improvement projects in Wikipedia.

Increasing contributions

In the human resource development literature, peer production communities are commonly referred to as "communities of practice" (CoP), a term from Lave and Wenger's [25] studies of professional training. Research on CoPs has studied participation barriers [12] and motivation [3], finding for example that many participants view their knowledge as a public good, so they have a moral obligation to share it.

Soliciting contributions from consumers of peer produced content is another way to increase activity. Halfaker et al. [15] nudged Wikipedia readers into submitting article feedback, which adds value as long as the system design makes it easy to weed out low quality contributions. A study of the Cyclopath bike-mapping community found that naturally occurring feedback on user behaviour could be used to improve suggested bike routes or add annotations to the map [28].

Intelligent task routing (ITR) uses a recommender system to match contributors with tasks related to their interest. It was first used to request contributions from users in a movie recommender system [7], finding that some personalised strategies were successful, but one was outperformed by the random baseline. ITR has also been implemented in an article recommender for Wikipedia called SuggestBot [8]; in this case, three personalised strategies were about four times more successful than a random baseline at eliciting article edits.

The social psychology literature has provided candidate interventions and appeals. Ling et al. [26] found that appealing to users' unique capabilities and giving them specific and challenging goals resulted in more contributions. In a followup study [32], Rashid et al. discovered that displaying the estimated value of a contribution had a positive impact. Further, identifying with the member group and the way a person viewed the member group also had a positive effect.

Other peer production communities focus on geographic information, known as Volunteered Geographic Information [13] (VGI). Increasing contributions to VGI communities has been studied in the context of Cyclopath [31]. Researchers found that users did considerably more work than explicitly requested, and that user familiarity with a given area strongly affected the type of work done.

Quality Improvement Projects in Wikipedia

Some of the improvement projects we examine in this paper have been studied previously in isolation. The Wikipedia Education Program (WEP) started in 2010 as the Public Policy Initiative (PPI)[7]. Lampe et al. [24] surveyed PPI participants, asking whether the project motivated them to continue contributing to Wikipedia after course completion. Students that reported actively participating and who were aware of Wikipedia's global reach were also more likely to say they would continue contributing. The APS Wikipedia Initiative is another project connected to the WEP. Farzan and Kraut [10] compared project participants with a cohort of similar Wikipedia contributors, finding that project participants added considerably more content, and that their content survived on par with that contributed by subject matter experts with PhDs. Lastly, the Wikimedia Foundation has published two reports on the quality of content added by students in the PPI and WEP projects [34, 43]. They used human assessment of a random selection of articles edited by project participants. The study of PPI found that the average article improved to an intermediate amount of quality, while the WEP study found a smaller increase in quality.

WikiProjects are self-organised groups of Wikipedia contributors interested in a specific topic area (e.g. WikiProject Military History) or type of work (e.g. WikiProject Wikify, which focuses on improving the layout and formatting of articles). These projects have been the focus of several research papers, studying for instance how they coordinate [20, 21], how they support collaboration [11], and how the diversity of group membership affects survival [6]. Some of the WikiProjects run a specific type of improvement project known as "Collaboration of the Week", a project that usually lasts a week or two and aims to improve a specific article or set of articles. These week-long collaborations were studied by Zhu, Kittur, and Kraut [45], who found that these article improvement goals strongly motivated project members to increase the amount of editing they did and that the effect also spilled over to other articles within the WikiProject's topic area.

Opportunity to integrate and expand

Research projects have used different methods and measures, making it difficult not only to compare one improvement project with another, but also to make comparisons within the same project. Take the Wikipedia Education Program, for example, which was studied by Farzan and Kraut [10] as well as the Wikimedia Foundation [34, 43]. Words added and word survival were the measures used in the former, while the latter used human assessment of the quality of a random

[7] `http://outreach.wikimedia.org/wiki/Public_Policy_Initiative`

selection of articles. Like this prior work, we also study the Education Program, but we apply a single form of assessment across this and several other improvement projects to enable cross-project comparisons.

We further extend the existing body of research by studying three improvement projects that have not been studied before:

- The WikiCup[8] is a competition for Wikipedia contributors started in 2007, whose purpose is to "encourage content improvement" and to be "just a bit of fun".
- Todays Article for Improvement[9] (TAFI) is a WikiProject started in mid-2012 that promotes an article per day, seeking to improve its quality.
- Wikipedia's Community Portal[10] is a page on Wikipedia that was created in early 2004 and has since featured an regularly updated list of articles in need of improvement.

In the next section, we present our descriptive framework before using it to describe the quality improvement projects we study. We then go through the datasets we gathered and our technique for measuring content quality in Wikipedia. This technique is then applied to our datasets, we report our findings, and discuss their impact. Lastly we consider some known limitations before final conclusions.

UNIFIED DESCRIPTIVE FRAMEWORK

In order to make comparisons across a diverse set of quality improvement projects, it is first necessary to identify a unified descriptive framework in which to understand these projects. We considered several candidates from the research literature, before settling on Preece's "Online Communities: Designing Usability and Supporting Sociability" [30]. Preece divides the social side of online communities into three components: **People**, **Purpose**, and **Policies**. We use each component as a major theme in our analysis and further explore the components as follows:

People: What *recruitment* method is used to find project participants, and is the work done by *individuals* or *groups*? Recruitment can either be *internal* – participants are already members of the community; *external* – participants are recruited from outside the community; or the project is open to anyone at all.

Purpose: The primary purpose for all the projects we examined was to improve the quality of Wikipedia. We are more interested in dimensions on which projects *differ* and therefore our specific analyses will focus on a project's secondary purpose, e.g. that students in the Education Program achieve academic course credit.

Policies: These comprise the governing structure for a project. Since Wikipedia itself has many policies and guidelines that influence all the projects we study, to avoid confusion, we use the term **structure** to describe the governing rules for individual projects.

[8] `https://en.wikipedia.org/wiki/Wikipedia:WikiCup`

[9] `https://en.wikipedia.org/wiki/Wikipedia:TAFI`

[10] `https://en.wikipedia.org/wiki/Wikipedia:Community_portal`

THE QUALITY IMPROVEMENT PROJECTS STUDIED

We sought a diverse set of improvement projects for our study. The five projects we study, and how they fit into our descriptive framework, are listed in Table 1. Below, we provide additional details:

- **Collaboration of the Week** (**CotW**). Some of the WikiProjects organise what is known as a **Collaboration of the Week**, where they focus on improving a specific article or a set of articles.

  As we see in Table 1 the **people** in CotW are *internal* as nearly all of them are already Wikipedia contributors and members of a specific WikiProject, and the work is done as a *group*. The collaboration's **purpose** is to achieve the group goal of improving a specific article or set of articles. Similarly the **structure** is a *group collaboration*. The vast majority of the collaborations last *a week or two*, on par with the name, but some last as long as a month.

- The **WikiCup** is a competition for Wikipedia contributors. Since 2009, the cup's organisation has been fairly stable, with four initial rounds followed by a final round, each round lasting approximately two months. There are comprehensive rules, and three judges award points. In each round contestants score points for achieving specific tasks. For example, contributing significantly to an article that successfully passes peer review for Featured Article status (the highest-quality Wikipedia article status) is awarded 100 points. In addition to the competitive aspects, it also is emphasised that the most important rule of the cup is "*just a bit of fun*" (emphasis theirs).

  As with CotW, WikiCup **people** are *internal* to Wikipedia, with contestants most likely already experienced members of the Wikipedia community. Work is done through (and assessed in terms of) *individual* effort. The **purpose** of the WikiCup is described on the cup pages. Of course, its primary purpose is to improve the encyclopedia, but as noted *scoring points* and *having fun* also are called out. Given the point scoring system, the cup **structure** involves *gamification* [9], and the duration is *months*.

- The **Wikipedia Education Program** (**WEP**) started as an organised effort connecting U.S. and Canadian university instructors and students with ambassadors from the Wikipedia community. The original intent of the program was for students to improve the content of public policy articles as part of class assignments. It has since expanded to other subject areas, countries, and languages. The Wikimedia Foundation says that there have been over 6,500 participants who have added "the equivalent of 45,000 printed pages of quality content"[11].

  The **people** in WEP are *external* to Wikipedia, specifically students at colleges and universities. In some courses, the students work individually on articles while in others they do group work, so we consider the project as having both. Since the work is done as part of college courses, we define the **purpose** of WEP to be *course credit*. Given the

[11] `https://outreach.wikimedia.org/w/index.php?title=Education/About&oldid=66258`

		CotW	WikiCup	WEP	Community Portal	TAFI
People	*Recruitment:*	Internal	Internal	External	Anyone	Anyone
	Individual or group work:	Group	Individual	Both	Individual	Group
Purpose	*Purpose:*	Group achievement	Scoring points, having fun	Course credit	Improve articles	Improve articles
Structure	*Structure:*	Group collaboration	Gamification	Academic coursework	None	None
	Duration:	Weeks	Months	Months	Hours	Day
	Study period:	2006–2009	2009–2013	2010–2013	Dec 2012	2012–2013
	Project size:	852	4,858	2,914	8,246	249

Table 1. The studied quality improvement projects. Abbreviations as follows: CotW: WikiProjects' Collaboration of the Week, WEP: Wikipedia Education Program, TAFI: Today's Article for Improvement. Project size is measured in number of articles.

shared context of post-secondary education and Wikipedia, the **structure** is *academic coursework*. Like the WikiCup, the duration of WEP courses is on the order of *months*, typically a U.S. semester of three to four months.

- The English Wikipedia's **Community Portal** (**CP**) serves several purposes, such as helping visitors learn what Wikipedia is about and how to do various Wikipedia tasks. However, what is relevant to our purposes is that it also features a list of articles that need improvement. The CP is easily accessed through a link in the menu on the left-hand side of any page on the English Wikipedia and is typically viewed about 10,000 times per day.

  While the **people** who visit the Community Portal are already on the Wikipedia site, they might not be members of the Wikipedia community (i.e., editors). Most Wikipedia articles do not require a registered account to be edited, and the Community Portal is a general call to action, which leads us to define the recruitment target as *anyone*. The CP does not feature any group collaboration or awareness mechanisms, so we regard it as *individual* work. The **purpose** of the list of articles that need improvement is simply article improvement. The CP provides no **structure**, and articles typically are promoted for one hour.

- **Today's Article for Improvement** (**TAFI**) is a WikiProject started in July 2012 with the goal of identifying "an undeveloped or underdeveloped article", which would then be promoted through various channels in Wikipedia. As of late May 2014 the project had 109 listed members.

  The **people** who participate in TAFI are recruited on Wikipedia through posts on project members' talk pages[12] and on the Community Portal, but also externally. For instance, some TAFI articles have been promoted on the official Wikipedia Twitter account. Thus, we define this project's recruitment target as *anyone*. Due to TAFI being organised by a WikiProject we see it as primary *group work*, but there likely also are individual efforts being made. There is no obvious secondary **purpose**, as the project is so clearly organised on improving a given article. No **structure** is provided, and the duration of TAFI is a single day.

[12]Every Wikipedia user has a talk page where they can be contacted.

Year	Our Data	Official Count	Our prop. (%)
2009	25	81	30.9
2010	53	135	39.3
2011	61	117	52.1
2012	51	111	45.9
2013	66	127	52.0
Total	256	571	46.4

Table 2. Overview of number of WikiCup participants per year in our dataset compared to the official number reported in the cup statistics.

DATASETS

Collaboration of the Week

We began with the collection of WikiProjects and articles studied by Zhu et al. We removed deleted articles, collaborations that targeted categories, and collaborations where it was unclear which article(s) they worked on. The result is a dataset of 852 articles spanning from 2006 to 2009.

WikiCup

Each WikiCup contestant has a page where they submit the work they have done for scoring review. We mined these pages for contestants in the cups from 2009 through 2013, as those cups have had the same format and a fairly stable scoring system. The result is a dataset with 256 contestants and 4,858 articles. This number of contestants is lower than the "official number" listed on the relevant WikiCup pages, Table 2 gives a yearly overview. We suspect this difference is because some users sign up but withdraw from the competition during a round or get disqualified. Therefore, we do not suspect this results in a distorted sample for our analysis.

Wikipedia Education Program

We mined three sources to gather a dataset covering 258 courses, 2,914 articles, and 2,870 students:

1. The U.S. and Canadian Education Program list of courses[13], which includes the Public Policy Initiative.

2. The Education Program extension's database, which covers the more recent courses in the program.

[13]`https://en.wikipedia.org/wiki/Wikipedia:Education_program/Courses`

3. The APS Wikipedia Initiative's Wikipedia page[14]. The APS Wikipedia Initiative is to some extent a separate project, but it still fits with the Education Program since some of the APS Wikipedia Initiative courses are included in the Education Program lists of courses.

We included data only from courses where individual students selected specific articles to work on, thus yielding an explicit record of the work done.

There also is an Indian Education Program that has worked on articles in the English Wikipedia. We did not include this project in our dataset for two reasons. First, the Wikimedia Foundation published a report[15] that described early contributions as "poor quality and/or ridden with copyright violations", and second, the remaining WEP courses form a fairly coherent group. There are now education programs in several countries and we plan to study these in future work.

The Community Portal

The list of articles that need improvement on the CP is updated automatically by a bot[16] roughly every hour. We mined the edit history of the Community Portal to gather a dataset of articles listed from December 4, 2012 to January 4, 2013. The bot updates 40 articles every time, and during the given time span 741 updates were made. Some articles were featured multiple times, resulting in a total of 8,246 unique articles.

Today's Article for Improvement

Our dataset of TAFI articles was collected from the project's archived schedule[17] as well as any article having the template "Former TAFI" applied to it. This resulted in a dataset containing 249 articles from July 2012 through December 2013.

Common properties of all datasets

For each article in each of the datasets, we gather the source (text and wiki markup) of the article at the start and end of every project. For the WikiCup and WEP, the end of the project is defined as the last edit by any project participant during the project. This is to ensure that we do not also capture additional work done by other editors. The remaining improvement projects are time-bound, e.g. the end of TAFI is the end of the day an article was selected.

We also gather data on the number of contributors working on each article between the start and end of each project. In the Education Program students assign themselves to specific articles, which provides an explicit mapping for us to use. For TAFI, CotW, and the WikiCup, we search the edit history of each article. We remove three categories of contributors: bots, because those are automated tools; those who were reverted by a bot or through common anti-vandal tools since they were likely vandalistic edits; and those who made reverts using common anti-vandal tools as that is maintenance work. The remaining set of contributors should be those who tried to make productive edits to an article.

[14] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Psychology/APS-Wikipedia_Initiative`

[15] `https://en.wikipedia.org/wiki/Wikipedia:India_Education_Program/Analysis/Quantitative_Analysis`

[16] Software robot, ref `https://en.wikipedia.org/wiki/Wikipedia:Bots`

[17] `https://en.wikipedia.org/wiki/Wikipedia:Today%27s_articles_for_improvement/Archives/Schedule`

Class:	Stub	Start	C	B	GA	A	FA
Quality:	*Low*						*High*

Table 3. English Wikipedia's seven assessment classes in order from lowest quality (left) to highest quality (right). Abbreviations: GA=Good Article, FA=Featured Article.

MEASURING PROJECT PERFORMANCE

The most important thing to measure about quality improvement projects is how much they improve the quality of the articles within their scope. This means we need a way to assess the quality of the articles in the datasets. There are several possible approaches to doing so, with the majority having appeared in the literature:

1. Using Wikipedia's own article quality assessments.
2. Gathering expert human assessment of randomly sampled articles (e.g. [34, 43]).
3. Crowdsourcing human assessment of a random sample of articles (e.g. [20]).
4. Using proxy measures for quality, e.g. words added and word survival (e.g. [10, 16]).
5. Leveraging machine learning techniques for predicting article quality (e.g. [42]).

Each one of these approaches comes with benefits and drawbacks. Wikipedia's own assessments are done by Wikipedia contributors using the seven-class scale shown in Table 3. This notion of article quality in Wikipedia has been shown to correspond well with existing notions of encyclopaedic quality [39]. However, because these assessments are done by people, there is a potential time lag between substantial changes to an article and its subsequent (re)assessment. As we are interested in measuring the immediate effect of article improvement work, the lag makes us unable to use the reassessments without further analysis.

The drawback of using experts to assess random samples is that it limits the number of samples, which can reduce statistical power. We come across a similar problem in one of our validation datasets, where 257 articles are not enough to tease out the effects we seek to understand (see Appendix A). While crowdsourcing assessments has been shown to be significantly correlated with Wikipedia's own assessments [20], the correlation ($r_s = 0.54$) also suggested disagreements. As we will show, we are able to produce higher correlations using a machine learning approach. Using proxy measures for quality would mean we would end up not capturing many features associated with article quality (e.g. the presence of references to sources or illustrative images). A machine learning approach enables measuring the entirety of the datasets, but will make prediction errors, requiring analysis of where prediction errors occur and how they affect overall results.

In this paper, we take a two-fold approach to quality estimation. Our primary focus is on a machine learning model,

	Est.	Std. Err.	P-val.
Intercept: $Stub\|Start$	-2.57	0.25	***
Intercept: $Start\|C$	-0.12	0.15	
Intercept: $C\|B$	1.30	0.16	***
Intercept: $B\|GA$	3.07	0.19	***
Intercept: $GA\|FA$	4.36	0.23	***
$log_2(n\_contributors)$	-0.51	0.06	***

Table 4. Ordinal Logistic Regression model coefficients for Collaboration of the Week. P-values: * < 0.001**

which we demonstrate can outperform published crowdsourcing approaches while at the same time providing assessments of a sufficient number of articles for our analyses. However, our model does have some error, and in order to validate our high-level results, we also repeat our experiments using a set of manual assessments from Wikipedia that we ensured were temporally valid. That is, the only assessments we considered in this second analysis were those that were applied to articles soon after a quality improvement project finished (so that the quality was not affected by subsequent edits). As we will show, we found substantial agreement between the model and the manual assessments. That is, our high-level model results hold across both means of assessing article quality.

The following section of the paper describes our results using the machine learning model to predict quality. In Appendix A, we describe the technical details of how the machine learning model is trained and evaluated, as well as demonstrate its strong performance relative to existing approaches. Appendix A also describes our validation of the results in the section that follows using manual assessments rather than the model's predictions and provides additional details about how we gathered these manual assessments.

RESULTS

This section focuses on three main findings. One relates to the **people** component of our framework, and a second to the **policies/structure** component. The third finding can be seen as relating either to **purpose** or **policies** depending on the improvement project's design. To complete our framework, we discuss this finding under the **purpose** component.

People

Result: More People, Less Quality

A fundamental question facing the designer of any effort to improve quality is: does it pay off to have contributors working individually on each artefact in the effort, or should they work in groups? Three of the studied projects have varying number of contributors per artefact, allowing us to investigate this question. Our results suggest that an *increased number of contributors per artefact* is associated with a *lower rate of increase in artefact quality*.

We examine the relationship between number of contributors and quality in the Collaboration of the Week (CotW), the Wikipedia Education Program (WEP), and the WikiCup. In all of these datasets, we have predicted the quality of each article at the start and end of the project using our quality machine learning model. We also calculated the number of contributors to each article during the project. The distribution

	Est.	Std. Err.	P-val.
Intercept: $Stub\|Start$	-2.94	0.10	***
Intercept: $Start\|C$	-1.26	0.07	***
Intercept: $C\|B$	0.44	0.07	***
Intercept: $B\|GA$	2.18	0.08	***
Intercept: $GA\|FA$	3.25	0.11	***
$from\_scratchTRUE$	0.47	0.08	***
$n\_contributors$	-0.10	0.03	**

Table 5. Ordinal Logistic Regression model coefficients for Wikipedia Education Program. P-values: ** < 0.01, * < 0.001**

	Est.	Std. Err.	P-val.
Intercept: $Stub\|Start$	-5.35	0.26	***
Intercept: $Start\|C$	-1.23	0.09	***
Intercept: $C\|B$	0.27	0.09	**
Intercept: $B\|GA$	0.43	0.09	***
Intercept: $GA\|FA$	3.15	0.11	***
$from\_scratchTRUE$	1.89	0.08	***
$log_2(n\_contributors)$	-0.63	0.04	***

Table 6. Ordinal Logistic Regression model coefficients for the WikiCup. P-values: ** < 0.01 * < 0.001**

of number of contributors is highly skewed in the CotW and WikiCup datasets. This is not uncommon for contributions to online communities. We therefore choose to log-transform these variables. The WEP dataset does not have the skewness issue. Group size in college classes is limited, so the most common size for WEP efforts is 2-5, and only a few outliers have more than 6 people.

To model the relationship between number of contributors and predicted quality we use an Ordinal Logistic Regression [27, 44] (OLR) with the assessment classes in the order shown in Table 3. We have a variable $n\_contributors$ for the number of contributors per artefact and add a binary variable $from\_scratch$ in the WEP and WikiCup dataset to control for articles that did not exist prior to the start of the project (thus having an unknown prior quality).

During our model building we also want to control for two additional issues: the *proportional odds assumption* and whether there is an *interaction effect* between our independent variables. The former is a fundamental assumption upon which OLRs are commonly built. In our case, it means the coefficients explaining the relationship between Stub-class ($P(Stub)$) and higher than Stub-class ($P(\geq Stub)$) also explain all other classes (e.g. $P(C)$ and $P(\geq C)$). We have verified that this assumption holds in all our OLR models. Second, we also verified that there is no interaction effect between our independent variables, which would have indicated that the strength of the effect of starting an article from scratch would be altered by the number of contributors to the article.

The results of our OLR models, one for each effort, are listed in Tables 4, 5, and 6. All predictors are statistically significant in all models. In the CotW model, the intercept (cutpoint) between Start and C-class is not significant. Because this cutpoint is only an estimate of the borderline between the two classes and the predictor's P-value < 0.001, this issue does not invalidate the model.

Across all three efforts the number of contributors has a negative sign indicating that larger numbers of contributors per artefact is associated with slower increase in quality. We also built additional models where we controlled for the quality at the start of the effort, to make sure that our model was not influenced by (for example) a larger proportion of articles starting from a certain quality level. Pre-effort quality was generally also a significant predictor in those models, but did not cancel out the effect of number of contributors. This means that consistently across these projects, an *increase in number of contributors per artefact* is connected with a *negative impact on the rate of quality increase*.

Discussion

We find it particularly interesting that the negative effect of additional contributors per artefact is consistent across all three projects, even though the nature of the "group" is different: in the WEP, participants are explicitly connected to an article, while in the other two projects we count all likely productive editors as participants. Wikipedia articles are of course open to anyone to edit, but WEP students are directed to work in a "sandbox", a personal space where they can draft an article before publishing it, as described in the template syllabi[18]. This usage of personal work spaces likely isolates many of the WEP articles from contributions from non-WEP contributors until they are published.

As we will see in the next section, WEP students seldom take articles above B-class quality, supporting the findings of the Wikimedia Foundation's studies on WEP quality [34, 43]. This could be due to a lack of experience with writing Wikipedia articles, but it could also be due to satisficing [35], they are doing just enough for a reasonable grade but nothing more. Groups of students might also be experiencing social loafing [18], e.g. that some of the group members are trying to free ride their way through the course while other members do the work. Future research on the WEP could try to tease these effects apart.

The groups of contributors in the WikiCup and CotW datasets are more implicit, and the extent to which participants in these efforts use sandboxes to edit articles before publication is unknown. It may be that additional contributors to those two efforts are not aware that they are taking part in an improvement project, which could alter their edit behaviour. These contributors may also differ in experience levels and engagement with the Wikipedia community, previous research has shown that power users in Wikipedia produce higher quality edits from early on [29]. Additional contributors could also be positive as long as only a few contributors are doing the majority of the work, as found by Kittur and Kraut [20], otherwise they just cause more maintenance overhead, similar to how adding people to late software development projects make them even later [4].

These results also beg the question of whether it is better for groups creating artefacts to work individually and sequentially. André et al. [2] found simultaneous work to be less effective than a sequential structure, but the effect was mitigated by assigning specific roles to participants. In Wikipedia there are few formal roles. Some users are promoted to become administrators, a role that is supposed to be janitorial and "not a big deal" [19] (yet research indicates it is an increasingly bigger deal [5]). Instead, users assume informal roles, which they may seek to use to their advantage in conflicts, for instance by questioning other contributors' expertise [22].

In order for a peer production community to be successful, there needs to be collaboration. These results suggests some degree of conflict between individual and group work, when does one approach benefit the community more than the other? We see investigations into how contributor roles, work organisation, conflict, coordination, and concentration of contributor effort affect artefact quality in improvement projects as a promising venue for future research.

	Stub	Start	C	B	GA	FA
NA	3.93	14.88	46.79	18.81	12.50	3.10
Stub	21.43	20.44	37.68	9.61	9.61	1.23
Start	0.95	25.79	44.94	17.25	8.07	3.01
C	0.00	1.49	54.29	25.00	12.87	6.34
B	0.00	1.21	16.92	65.86	6.95	9.06
GA	0.00	0.00	10.58	11.54	61.54	16.35
FA	0.00	0.00	3.08	6.15	7.69	83.08

Table 7. Prior (rows) and post (columns) predicted quality for the Wikipedia Education Program. Proportions are relative to prior quality (rows). NA=Article did not exist prior to start of a course.

	Stub	Start	C	B	GA	FA
NA	0.90	34.11	37.95	3.60	21.46	1.98
Stub	1.18	28.10	40.26	2.09	26.67	1.70
Start	0.00	24.25	17.55	2.08	51.96	4.16
C	0.00	1.18	47.14	3.16	39.05	9.47
B	0.00	0.00	5.74	36.89	47.54	9.84
GA	0.00	0.19	1.23	0.85	89.26	8.48
FA	0.00	0.33	0.33	0.66	7.95	90.73

Table 8. Prior (rows) and post (columns) predicted quality for the WikiCup. Proportions are relative to prior quality (rows). NA=Article did not exist prior to start of a cup round.

Purpose

Result: New Artefacts, Higher Quality

Is it more effective to have participants in a quality improvement project create new artefacts or work on improving existing ones? Two of our efforts, the WikiCup and the Wikipedia Education Program include both types of work. Across both of these, our results indicate that artefacts created from scratch end with a higher final quality.

To investigate this effect, we first look at what level of quality articles reach at the end of a project. Table 7 (WEP) and Table 8 (WikiCup) show the relationship between predicted quality at the start of an improvement project (rows) and at the end (columns), where Good Article is abbreviated "GA" and Featured Article "FA". For the WikiCup, the end of the project

[18] `http://outreach.wikimedia.org/w/index.php?title=Education/Syllabi&oldid=70162`

[19] `https://en.wikipedia.org/wiki/Wikipedia:DEAL#History`

	Stub	Start	C	B	GA	FA
Stub	29.46	40.31	17.83	6.20	4.65	1.55
Start	1.74	46.52	27.39	18.26	6.09	0.00
C	0.00	0.65	69.48	17.53	10.39	1.95
B	0.00	1.45	14.49	72.83	6.52	4.71
GA	0.00	0.00	4.88	2.44	80.49	12.20
FA	0.00	0.00	0.00	9.09	9.09	81.82

Table 9. Prior (rows) and post (columns) predicted quality for Collaboration of the Week. Proportions are relative to prior quality (rows).

is the last edit done by a specific participant on the article that participant submits for scoring, and for the WEP the end is the last edit done by any student assigned to a specific article. For convenience, we have also included the same type of table for the Collaboration of the Week (Table 9) but note that in that project no articles were created from scratch.

In the WEP (Table 7), 65.6% of the articles started from scratch (the "NA" row) reach an intermediate level of quality (C- or B-class). This is not the case for the WikiCup (Table 8), where instead more than one fifth of every new article is predicted as Good Article (GA) or Featured Article (FA) status. We can also see that to a certain degree in the WikiCup, and to a much larger degree in the WEP, many articles do not improve enough to change their predicted quality class.

More generally, our OLR models in Tables 5 and 6 show that the $from\_scratch$ variable is a significant predictor with a positive relationship to end quality. This suggests that in both the WikiCup and WEP projects, *new artefacts have higher end quality* compared to existing artefacts.

Discussion

Here again we have found an effect that is consistent across vastly different improvement projects. As previously noted, many WEP articles are likely isolated from contributions from non-WEP editors due to the extensive use of sandboxes. In contrast, the WikiCup has an "In the news" category for articles that are featured in that section on the English Wikipedia's front page, with the contestant scoring 10 points for each article featured. This will likely lead to the cup containing some breaking news articles [19], newly created articles where the particularly high interest and resulting traffic could lead to quicker improvements in quality.

The result also is interesting because both projects have a long duration, namely months. With that amount of time available, one would not expect there to be a significant difference in quality improvement between new and existing articles, particularly one in favour of new articles. Producing high-quality Wikipedia articles requires access to resources, for instance sources for claims and illustrative images. For some types of content these might be more difficult to find, particularly using online resources, and in the case of existing artefacts resources might already have been exhausted. The lack of online sources could to some extent explain the WikiCup result where participants might strongly prefer them, but it seems unlikely to explain the WEP result, since students should have access to good library resources.

Existing artefacts are also more likely to have some contributors monitoring them. Research on Wikipedia has shown that editors assume ownership of content [16, 41] although Wikipedia's own policy states noone owns an article[20]. This type of territoriality also occurs outside of Wikipedia, expert contributors to a museum tagging system were found to more strongly express ownership of content than novices [40]. When participants in a quality improvement project try to make changes, territoriality by existing contributors is likely a barrier to entry, resulting in reduced quality gain through coordination overhead.

It is also not obvious that peer production communities should always focus on creating new artefacts. If the community already has good coverage (e.g. a large number of articles), it would perhaps instead benefit the most if work was concentrated on improving existing artefacts. Community managers could combine the understanding of this trade-off between coverage and quality with information on audience attention to guide contributions to the areas where they are most needed in order to ensure the community's resources are utilised most efficiently.

Perhaps people work differently if they start with a blank slate than if they have to modify an existing piece of work. In their study on the effect of seeding wikis with content, Solomon and Wash [36] found that not seeding led to significantly more content added, while those who started with seeded content would instead use that as a model. We do not know to what extent this finding also is present in the work WikiCup and WEP participants do on existing articles. There is an opportunity here for both qualitative analysis of live data as well as lab studies to understand the effects that are in play and how to most efficiently produce high quality artefacts.

We also found interesting differences between improvement projects in the patterns of change in predicted quality. The WikiCup results (Table 8) show that few articles move into the B-class. Instead, the cup participants push articles upwards to GA/FA status, a behaviour we interpret to be clearly in line with the cup's incentive mechanism. Successful Good Article nominations score 30 points and Featured Articles 100 points, while getting an article only to B-class scores zero. This is similar to how badges steer user behaviour in Q&A systems [1]: when users have nearly reached a badge threshold, they will modify their behaviour to achieve the badge as quickly as possible.

The results for WEP (Table 7) and the Collaboration of the Week (Table 9) show that for many articles quality does not appear to improve. The CotW's short duration, usually a week or two can explain the effect in that project. Most improvements in CotW occur in low quality articles, confirming Zhu et al.'s description of those articles being the typical collaboration targets [45]. That the Education Program also to a large extent leads to improvements that appear to not substantially change the article quality is more concerning. Students in the program have more time available to affect change, thus we wonder if they are struggling with learning how to write

[20] `http://en.wikipedia.org/wiki/Wikipedia:Ownership_of_articles`

articles in the context of Wikipedia, for instance how to correctly source content with footnotes and citation templates.

Policies/Structure

Result: Structure is Required

Two of the improvement projects we study, Today's Article for Improvement (TAFI) and Wikipedia's Community Portal, do not have a well-defined structure. For example, they are open to anyone, have a very general purpose, and lack a clear incentive mechanism. Our results indicate that unlike the other projects, neither TAFI nor the Community Portal is particularly successful at improving artefact quality.

First, let us investigate the TAFI project. We predicted the quality of each article at the beginning and end of the effort, in this case the day the article was promoted for improvement. Only 9 out of 249 articles (3.6%) saw an improvement in predicted class, and of those all but one moved up a single class, the exception being a Start-class article improving to B-class.

Is the problem lack of participation? TAFI started in mid-2012 and at the end of 2013 the project's member list contained 103 usernames. Still, of the 249 articles in the TAFI dataset 56.2% had no contributors during the day of the effort. We investigated whether the degree of participation changed over the course of the project and found that in the first three months, 1 out of 16 articles saw no contributors, while in the last three months it was 33 out of 47. This is a statistically significant difference (Fisher exact count test $p < 0.001$); the project has seen a significant decline in participation as its membership has increased.

We found a similar problem with participation in the Community Portal. Our dataset covers Dec 4, 2012 to Jan 4, 2013 during which time the portal, according to data from the Wikimedia Foundation[21], saw 314,534 views, for a daily average of 10,146 views. One would hypothesise that these views would directly affect listed articles as visitors to the portal follow links to edit them. To investigate this, we calculated average views/day prior to being listed for the portal articles and removed those that were listed twice on the same day due to our view data having a granularity of one day. We sorted the articles into buckets based on average views/day, using exponential buckets since article popularity follows a power-law distribution. Lastly we calculated views on the day of listing, as well as average views/day up to 14 days after.

Articles are typically listed for only one hour, so one would expect the portal to affect article views less as popularity increases. This is also seen in Table 10, which shows an excerpt of the view results. The remaining part of the table (up to x $\geq$ 16,384 views/day) is left out for brevity as the trend of a decrease in views on the listed day as well as in the period after continues. Based on the results in Table 10 it seems clear that few views appear to come from the Community Portal.

Not surprisingly, since including an article on the Community Portal did not increase how much it was viewed, it also didn't increase participation. We selected portal articles which had no edits in the two weeks prior to being listed, because those articles are most likely getting contributions from the portal. Out of 4,410 articles only eight of them were edited during the time they were listed. This extends data from the Wikimedia Foundation during a redesign of the portal in late 2012, where 220,000 portal views led to 46 saved edits [22]. Since the portal does not lead to participation, there obviously can be no improvements in quality. Therefore it is not an example of a successful improvement project.

Discussion

Both of the unstructured projects studied were largely unsuccessful. The short duration of these projects, an hour in the case of the Community Portal and a day for the TAFI project, might be posited as the explanation for the lack of success. However, the Community Portal is easily accessible from the left-hand menu of any Wikipedia page, and as we saw exposes a lot of readers to its call to action. In our related work we referenced several successful projects with a much similar approach: Cosley et al. [7] suggested edits of movie data on a movie recommender site; a general call to action solicited contributions in the Cyclopath geo-wiki [31]; Halfaker et al. [15] asked Wikipedia readers to submit article feedback. In all three cases more structure and guidance was supplied when necessary, for instance Cosley et al. had a form for inputting data, and the Cyclopath experiment provided volunteers with clear instructions for the work needed.

Comparing TAFI and the Community Portal to the other projects, we also see that these two unsuccessful projects lack a clear purpose, perhaps it is unclear to potential participants what the benefit is to both them and the encyclopaedia. In contrast, many of the WEP courses aim to extend Wikipedia's content in areas where it is lacking (e.g. public policy or psychology), and the WikiCup's scoring system appears to steer participant behaviour, as seen in their movement of articles to the higher quality classes to score points. Neither TAFI nor the Community Portal implements similar incentive mechanisms. Where our initial investigation has pointed to a lack of participation, future work could look at how much structure and what kind of incentive mechanisms are needed to trigger increased participation to cross the border into a successful improvement project.

LIMITATIONS AND NEXT STEPS

This research has several known limitations. First, while the English Wikipedia is the largest peer production community in existence, results from this community might not generalise. For example, a Q&A system like Stack Overflow is also a peer production community with some wiki-like features. Research to determine to what extent our findings also are present there (or in other peer production communities) would be valuable.

Second, our analysis uses Wikipedia's own assessment classes. Wikipedia's notion of article quality has been shown to correspond well with existing notions of encyclopaedic quality [39]. In our analysis of prediction errors, we discovered that in some cases Wikipedia contributors failed to apply

[21] `http://dumps.wikimedia.org/other/pagecounts-raw/`

[22] `https://meta.wikimedia.org/wiki/Research:Community_portal_redesign/Opentask`

Views/Day Bucket	Prior mean views	Listed day gain/loss (%)	Post gain/loss (%)
$0 \leq x < 2$	1.6	126.08	80.34
$2 \leq x < 4$	3.1	33.12	12.10
$4 \leq x < 8$	5.6	6.74	-4.36
$8 \leq x < 16$	11.2	-3.68	-8.50

Table 10. Excerpt of view statistics for articles listed on Wikipedia's Community Portal. Articles are placed in buckets based on prior mean views.

the assessment criteria correctly, leading to articles being assessed into a lower class. This suggests that there is room for improvement in the understanding of how Wikipedia contributors apply the assessment criteria, as well as how these correspond to assessment of quality by non-Wikipedians, and we plan future work in this area.

This paper brings together a diverse set of improvement projects, which means we must also consider limitations imposed by them. There is likely a clear difference in skill levels between some of the efforts. Contestants in the WikiCup and WikiProject members participating in the Collaboration of the Week are probably skilled members of the Wikipedia community, while students in the Education Program have little prior experience with writing for Wikipedia. One way to control for this would be to introduce measures of tenure, for instance the number of edits a contributor has or the amount of time since account registration.

We are also limited by how we define a contributor to a specific article. In the Education Program we use the course pages' explicit definition of which students worked on a specific article, and in the WikiCup we use contestants' submission pages to track which articles they worked on as part of the cup. In the other efforts, we instead use an implicit method of defining participants. This method could potentially be improved by algorithmic content analysis, for instance to account for different categories of contributors (e.g. newly registered users, users without an account, etc).

CONCLUSION

We have studied five diverse quality improvement projects in the English Wikipedia. Our findings suggest three important implications for design:

1. **People:** Increasing number of contributors per artefact is associated with slower increase in quality. Consideration should be given to when working individually can be more effective than group work.
2. **Purpose:** Artefacts created during the improvement project are connected to a higher quality level than existing artefacts worked on during the project. There may even be cases where deleting an old artefact to start over is preferred, although more research is needed.
3. **Policies/Structure:** Unstructured efforts are less likely to succeed. Our results suggest that new efforts should provide a carefully designed socio-technical structure, for instance through incentive mechanisms appropriate for the desired work and the knowledge level of the participants.

To analyse a diverse set of quality improvement projects, we used Preece's three components of online communities (outlined in bold above) as building blocks for an analytic framework. We developed a classifier for assessing Wikipedia article quality and verified its performance, enabling us to bring several existing threads of research together, while at the same time extending the variety of improvement projects studied. This research provides researchers and designers with the knowledge to design and evaluate more effective quality improvement projects in the future.

ACKNOWLEDGEMENTS

We would like to thank Haiyi Zhu and colleagues for access to their Collaboration of the Week dataset, our GroupLens colleagues for their support, the reviewers for their helpful assistance in improving the quality of this paper, the Wikimedia Foundation for facilitating access to Wikipedia data, and all Wikipedia contributors for creating a great encyclopaedia for us to study. This work has been funded in part by the National Science Foundation (grants IIS-0808692, IIS-0968483, and IIS-1111201) and a Yahoo! ACE Award.

APPENDIX A: CLASSIFIER TRAINING AND EVALUATION

To create a classifier that can assess the quality of Wikipedia articles, we built upon our previous work [42] where a Random Forest (RF) classifier was used to predict which of Wikipedia's seven assessment classes (shown in Table 3) to which an article belongs. In that study we found that the RF classifier had the best overall performance, and we therefore chose to use an RF classifier as our starting point. We improve the classifier's performance through four steps:

1. A much larger dataset (N=29,828), which requires us to address the class imbalance problem imposed by Wikipedia's low number of A-class articles.
2. A larger set of quality features extracted from each article.
3. Each feature is tested six times using 10-fold cross-validation to determine how each feature most strongly relates to article quality (raw metric, log-transformed, and four variants of proportions relative to article length).
4. Classifier parameter tuning (again using 10-fold cross validation) to determine forest size, the number of features to use in each tree split, and terminating node size.

There is no gold-standard dataset on which to train a classifier for this task. To gather a suitable set of candidate articles we copied the behaviour of WP 1.0 Bot[23], the software robot that gathers statistics on Wikipedia article assessments. Using Wikipedia's category system to find articles in a specific assessment class we collected 29,828 article assessments, 5,000 from each class with two exceptions: the Featured Article (FA) class had 4,062 articles at that time, and we only found

[23] `https://en.wikipedia.org/wiki/User:WP_1.0_bot`

766 A-class articles. Official statistics listed 1,279 A-class articles and the discrepancy is likely due to duplicates.

The low number of A-class articles creates what is known as a *class imbalance problem* [17]. Random Forest classifiers require reasonably balanced classes, so without remedial action, this would result in poor classifier performance on A-class articles. Typical approaches are oversampling the smaller class, or undersampling the larger classes. We tested both of these approaches and found that they led to lower classifier performance. The 766 articles accounted for only 0.018% of the total number of articles in the English Wikipedia at that time, so, statistically speaking, this article class simply is not used in the encyclopedia. Given the low usage of this class, the probability of an article in our datasets belonging to it is very low, which means that removing it does not significantly impact our study. Therefore, we decided to ignore A-class articles altogether, and we confirmed this significantly increased classifier performance.

WikiProjects "claim" – and thus assess – articles, and multiple projects can claim the same article (e.g., the Barack Obama article is claimed by 14 projects). How do we select an assessment class for an article if different projects disagree on its assessment? We looked at two methods – (1) choose the **highest** class, (2) choose the **majority** class – and found that these two methods disagreed on only 150 out of 29,828 articles (Cohen's Kappa = 0.967 with two raters, p-value $\ll 0.001$). Therefore, we chose to use the highest assessment class as an article's correct class.

For each article in our training dataset, we went through the article's assessment history to find the point in time where it first belonged to a given class. If that revision is not available, for instance revisions sometimes get deleted due to copyright issues, we used the first available more recent revision. We then retrieved the revision content (text and wiki markup) and extracted the following 11 features:

1. article length in bytes (log-transformed)
2. number of references (log-transformed)
3. number of links to other articles (log-transformed)
4. number of citation templates
5. number of non-citation templates (log-transformed)
6. number of categories linked in the article text
7. number of images / article length
8. information noise score (as defined by Stvilia et al. [38])
9. has an infobox template (binary variable)
10. number of level 2 section headings
11. number of level 3+ section headings

In order to verify classifier performance on this dataset, we chose to split the dataset using random selection to get a training dataset (80%) and a test dataset (20%). Using 10-fold cross-validation on the training set we validated our features and identified optimal classifier parameters. A forest with 501 trees and terminating node size 8 showed the best performance. Training a classifier on the entire training dataset and validating its performance on the test set results in the confusion matrix shown in Table 11.

	FA	**GA**	**B**	**C**	**Start**	**Stub**	**N**
FA	546	167	47	9	1	0	770
GA	252	655	54	83	5	0	1049
B	81	151	374	261	129	11	1007
C	25	128	201	471	189	18	1032
Start	1	12	71	201	600	138	1024
Stub	0	0	0	14	166	818	998

Table 11. Confusion matrix for the Random Forest classifier predicting articles in the test set. Rows (italic) are true (assessed) class, columns (bold) are predicted class. Last column (N) is the total number of articles in each class.

	CotW		**WEP**		**WikiCup**	
Distance	**N**	**%**	**N**	**%**	**N**	**%**
5					1	0.1
4			2	0.8		
3	2	3.5	7	2.7	26	2.4
2	8	14.0	38	14.8	40	3.7
1	12	21.1	104	40.5	223	20.8
0	27	47.4	96	37.4	699	65.1
-1	7	12.3	7	2.7	49	4.6
-2	1	1.8	3	1.2	31	2.9
-3					4	0.4
Total	57	100.0	257	100.0	1,073	100.0

Table 12. Prediction errors by distance between reassessed and predicted class for the Collaboration of the Week (CotW), Wikipedia Education Program (WEP), and Wikicup. Positive error means the prediction was a higher quality class.

The difference in number of articles per class in Table 11 is due to fewer Featured Articles (FA) and the random selection. We see that the overall error rate is 41.08%. Similarly as in our previous work, the classifier is often off by one class. If we allow one class leeway the error rate drops to 10.5%. The classifier also often errs on the high side, for instance more Start-class articles are predicted as C than Stub.

While the performance of the classifier on the test dataset is promising, we also wanted to verify its performance on articles after completion of a quality improvement project so as to understand its performance on data more specifically associated with the goals of this research project. Three datasets of articles that were reassessed after a project's completion were gathered, one each from the Collaboration of the Week (518 articles), the Wikipedia Education Program (987 articles), and the WikiCup (1,617 articles).

Many of these reassessments suffer from the time lag described in the section "Measuring Project Performance". For instance in the CotW dataset the median time to reassessment is 157.6 days. In the intervening time the article may have gone through substantial changes. We therefore restricted these datasets to reassessments that occurred within 10 edits, and where the article has changed by less than 100 bytes. When checking some of our prediction errors described below, we also confirmed that this edit/size limitation led to articles only going through minor changes, e.g. copy edits. After applying this limitation we were left with 57 CotW articles, 257 WEP articles, and 1,073 WikiCup articles.

Dataset	Reassessments	Predictions
CotW	N collaborators negatively associated with quality. Significant ($p < 0.01$).	N collaborators negatively associated with quality. Significant ($p < 0.01$).
WEP	Not statistically significant.	Not statistically significant.
WikiCup	N collaborators negatively associated with quality. New articles positively associated with quality. Significant ($p < 0.001$).	N collaborators negatively associated with quality. New articles positively associated with quality. Significant ($p < 0.001$)

Table 13. Comparison between Ordinal Logistic Regression models built on article reassessments and predictions for each reassessment dataset.

For each article, we then predicted their quality class as described earlier in order to enable the comparison of predictions against human assessments post quality improvement. As we were interested in learning specifically to what extent the classifier makes prediction errors, and when it does, how severe these errors are, we chose to measure the error as the distance between predicted and assessed class along the ordinal scale listed in Table 3 (e.g. a B-class article predicted to be Start has an error of -2). We then summed these errors across all classes. The distribution of prediction errors for each reassessment dataset is shown in Table 12.

For the WikiCup, the classifier shows stronger performance than on our test set, while the performance is less for the other two. The large proportion of one-class errors in the WEP dataset led us to investigate further, finding that the majority of the errors come from articles reassessed as C-class (23 articles) and Start-class (64 articles). A random sample of 23 Start-class articles and all 23 B-class articles were selected and verified that they had all gone through only minor changes (e.g. link fixes or minor copy-editing). The English Wikipedia's criteria for Start-class assessment states in part that the article "most notably, lacks adequate reliable sources." Inspection of the Start-class articles by an expert Wikipedia contributor indicated that the vast majority (20 articles) appeared to have several, if not many, reliable sources, suggesting that this subset of articles were not correctly reassessed, an issue that is also discussed in our "Limitations" section and that opens to future work.

The classifier's predictions are strongly correlated with Wikipedia's own article assessments, more so than using a crowdsourcing approach. This is the case across all four evaluation datasets: the test set ($r_s = 0.86$), CotW ($r_s = 0.57$), WEP ($r_s = 0.58$), and the WikiCup ($r_s = 0.82$). In all these cases we have a higher correlation than what was reported when crowdsourcing was used ($r_s = 0.54$) [20].

Using the post-improvement reassessments to perform the same analysis that forms the basis for this paper gives the same results. To determine this we built two Ordinal Logistic Regressions for each of the three reassessment datasets, one each using the reassessment and the prediction as the dependent variable. This set of models can then be checked for agreement and the results are listed in Table 13.

Aside from the lack of statistical significance for WEP, we see in Table 13 that in all cases the pair of models agree with each other. When significant, number of contributors is negatively associated with post-improvement quality, and creating a new artefact is positively associated with quality. Based on the classifier's performance as established by this appendix and its agreement with post-improvement manual assessments, we therefore conclude that our classifier-based results hold when we are analysing the entire datasets, which also allows us to gain statistical significance for WEP.

REFERENCES

1. Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. Steering user behavior with badges. In *Proc. of WWW* (2013), 95–106.
2. André, P., Kraut, R. E., and Kittur, A. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. In *Proc. of CHI* (2014), 139–148.
3. Ardichvili, A. Learning and Knowledge Sharing in Virtual Communities of Practice: Motivators, Barriers, and Enablers. *Advances in Developing Human Resources 10*, 4 (2008), 541–554.
4. Brooks, F. P. *The Mythical Man-Month*, vol. 1995. Addison-Wesley Reading, 1975.
5. Burke, M., and Kraut, R. Mopping Up: Modeling Wikipedia Promotion Decisions. In *Proc. of CSCW* (2008), 27–36.
6. Chen, J., Ren, Y., and Riedl, J. The Effects of Diversity on Group Productivity and Member Withdrawal in Online Volunteer Groups. In *Proc. of CHI* (2010), 821–830.
7. Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. Using Intelligent Task Routing and Contribution Review to Help Communities Build Artifacts of Lasting Value. In *Proc. of CHI*, CHI '06 (2006), 1037–1046.
8. Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. In *Proc. IUI* (2007), 32–41.
9. Deterding, S., Dixon, D., Khaled, R., and Nacke, L. From game design elements to gamefulness: Defining "gamification". In *Proc. of Mindtrek* (2011), 9–15.
10. Farzan, R., and Kraut, R. E. Wikipedia classroom experiment: Bidirectional benefits of students' engagement in online production communities. In *Proc. of CHI* (2013).
11. Forte, A., Kittur, N., Larco, V., Zhu, H., Bruckman, A., and Kraut, R. E. Coordination and Beyond: Social Functions of Groups in Open Content Production. In *Proc. of CSCW* (2012), 417–426.

12. Garfield, S. 10 reasons why people don't share their knowledge. *KM Review 9*, 2 (2006), 10–11.

13. Goodchild, M. Citizens as sensors: the world of volunteered geography. *GeoJournal 69*, 4 (2007), 211–221.

14. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design 37*, 4 (2010), 682.

15. Halfaker, A., Keyes, O., and Taraborelli, D. Making Peripheral Participation Legitimate: Reader engagement experiments in Wikipedia. In *Proc. of CSCW* (2013), 849–860.

16. Halfaker, A., Kittur, A., Kraut, R., and Riedl, J. A jury of your peers: quality, experience and ownership in Wikipedia. In *Proc. WikiSym* (2009), 15:1–15:10.

17. Japkowicz, N., and Stephen, S. The class imbalance problem: A systematic study. *Intelligent data analysis 6*, 5 (2002), 429–449.

18. Karau, S. J., and Williams, K. D. Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology 65*, 4 (1993), 681.

19. Keegan, B., Gergle, D., and Contractor, N. Hot off the wiki: Structures and dynamics of wikipedias coverage of breaking news events. *American Behavioral Scientist 57*, 5 (2013), 595–622.

20. Kittur, A., and Kraut, R. E. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *Proc. CSCW* (2008).

21. Kittur, A., Pendleton, B., and Kraut, R. E. Herding the Cats: The Influence of Groups in Coordinating Peer Production. In *Proc. of WikiSym* (2009), 7:1–7:9.

22. Kriplean, T., Beschastnikh, I., McDonald, D. W., and Golder, S. A. Community, Consensus, Coercion, Control: Cs*W or How Policy Mediates Mass Participation. In *Proc. of GROUP* (2007), 167–176.

23. Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proc. of WikiSym* (2011), 1–10.

24. Lampe, C., Obar, J., Ozkaya, E., Zube, P., and Velasquez, A. Classroom wikipedia participation effects on future intentions to contribute. In *Proc. of CSCW* (2012), 403–406.

25. Lave, J., and Wenger, E. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

26. Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M., et al. Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication 10*, 4 (2005), 00–00.

27. Liu, I., and Agresti, A. The analysis of ordered categorical data: An overview and a survey of recent developments. *Test 14*, 1 (2005), 1–73.

28. Masli, M., and Terveen, L. G. Leveraging the Contributory Potential of User Feedback. In *Proc. of CSCW*, CSCW '14, ACM (New York, NY, USA, 2014), 956–966.

29. Panciera, K., Halfaker, A., and Terveen, L. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proc. GROUP* (2009), 51–60.

30. Preece, J. *Online communities: Designing Usability and Supporting Socialbilty*. John Wiley & Sons, Inc., 2000.

31. Priedhorsky, R., Masli, M., and Terveen, L. Eliciting and Focusing Geographic Volunteer Work. In *Proc. of CSCW* (2010), 61–70.

32. Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, R., and Riedl, J. Motivating Participation by Displaying the Value of Contribution. In *Proc. of CHI*, CHI '06 (2006), 955–958.

33. Reagle, J., and Rhue, L. Gender Bias in Wikipedia and Britannica. *International Journal of Communication 5*, 0 (2011).

34. Roth, A. Student Contributions to Wikipedia. `https://outreach.wikimedia.org/wiki/Student_Contributions_to_Wikipedia`. Retrieved June 2, 2014.

35. Simon, H. A. Rational choice and the structure of the environment. *Psychological review 63*, 2 (1956), 129.

36. Solomon, J., and Wash, R. Bootstrapping wikis: developing critical mass in a fledgling community by seeding content. In *Proc. CSCW* (2012), 261–264.

37. Stephens, M. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* (2013), 1–16. 00001.

38. Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. Assessing information quality of a community-based encyclopedia. In *Proc. ICIQ* (2005), 442–454.

39. Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. Information quality work organization in Wikipedia. *J. Am. Soc. Inf. Sci. Technol. 59* (April 2008), 983–1001.

40. Thom-Santelli, J., Cosley, D., and Gay, G. What Do You Know?: Experts, Novices and Territoriality in Collaborative Systems. In *Proc. of CHI* (2010), 1685–1694.

41. Thom-Santelli, J., Cosley, D. R., and Gay, G. What's Mine is Mine: Territoriality in Collaborative Authoring. In *Proc. of CHI* (2009), 1481–1484.

42. Warncke-Wang, M., Cosley, D., and Riedl, J. Tell me more: An actionable quality model for wikipedia. In *Proc. of WikiSym* (2013), 8:1–8:10.

43. WMF. Spring 2012 United States and Canada student article quality research results. **https://en.wikipedia.org/wiki/Wikipedia:Ambassadors/Research/Article_quality/Results**. Retrieved June 2, 2014.

44. Yee, T. W. The vgam package for categorical data analysis. *Journal of Statistical Software 32*, 10 (2010), 1–34.

45. Zhu, H., Kraut, R., and Kittur, A. Organizing without formal organization: Group identification, goal setting and social modeling in directing online production. In *Proc. of CSCW* (2012), 935–944.