The Effects of Group Composition on Decision Quality in a Social Production Community

Shyong (Tony) K. Lam[1] Jawed Karim[2] John Riedl[1]

[1]GroupLens Research
Department of Computer Science and Engineering
University of Minnesota
{lam,riedl}@cs.umn.edu

[2]Computer Science Department
Stanford University
jawed@cs.stanford.edu

ABSTRACT

Online social production communities allow efficient construction of valuable and high-quality information sources. To be successful, community members must be effective at collaboration, including making collective decisions in the presence of disagreement. We examined over 100,000 decisions made by small working groups in Wikipedia, and analyzed how decision quality in these online groups is influenced by four group composition factors: the size of the group, how members were invited to the group, the prior experience of group members, and apparent bias shown by the group administrator. Our findings lead us to recommendations for designers of social production communities.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software—*Information networks*; H.5.3 [**Information Systems**]: Group and Organization Interfaces—*computer-supported collaborative work*

General Terms

Human Factors, Measurement

Keywords

Wikipedia, decision making, small groups, collaboration

1. INTRODUCTION

Over the past decade, online social production communities have become an increasingly viable and popular way to create high-quality sources of information without requiring the services of professionals or experts. Question answering systems such as *Yahoo! Answers* and *StackOverflow* allow people to directly help each other solve problems and find answers to previously asked questions. Social news sites like *Digg* rely on community submissions and social voting to produce interesting news feeds. *Wikipedia*, which is written and maintained by its community, has become one of the world's most popular sources of information.

A social production community relies on collaboration among its members to thrive. Disagreement and conflict will inevitably arise, and the community must make decisions about how to resolve conflict and move forward. Effective decision-making and conflict resolution processes are essential to a healthy community. Flawed processes may lead to poor decisions, which are costly to address—not only can they increase coordination costs and process losses, they can also alienate users and cause them to leave.

GROUP'10, November 7–10, 2010, Sanibel Island, Florida, USA.

1.1 Contributions

We explore group decision-making process in the context of social production communities. We analyze over 100,000 content decisions made by small working groups in Wikipedia, and study how four group composition factors affect decision quality. Our results lead us to a number of recommendations and implications for the design of online social production communities.

1.2 Related Work and Research Questions

Group decision making is a rich area of research that has been studied extensively in multiple disciplines, including social psychology, economics, and political science [9, 13]. One limitation of the existing literature is that much of it focuses on group composition factors that affect performance in face-to-face settings. A goal of the current work is to learn how these factors apply in computer-mediated communication (CMC) settings where group members are in an environment that lacks nonverbal and paraverbal cues, often working asynchronously and with anonymous or pseudonymous peers. Because of these differences, we cannot assume that the findings from offline groups will apply in an online context.

There have been been numerous comparisons of group performance between groups that use CMC and those that use face-to-face communication (e.g., [11, 16]), but many are limited to studying the broad performance differences between groups in offline and online environments. A meta-analysis by Baltes et al. examined 27 such studies and found that CMC groups generally underperformed face-to-face groups [1]. CMC groups took longer to make decisions, made worse decisions, and had lower member satisfaction. The meta-analysis found several factors that influenced the effectiveness of CMC groups, including anonymity, group size, and task type. We seek to expand on this knowledge and find ways for designers of online social production communities to improve their decision-making processes.

We now review the literature on group decision making and conflict resolution, focusing on and highlighting several group composition factors that influence decision-making acuity in face-to-face settings. These factors will become the basis for our research questions in the current work.

Social psychologists have found that group size affects the dynamics of conflict resolution processes in small groups. People in

larger groups are prone to escalating conflict and are less likely to cooperate with one another [13]. Larger groups also suffer from process losses, which may reduce efficiency and performance [17]. On the other hand, in *The Wisdom of Crowds*, Surowiecki suggests that in many cases, large and diverse groups can make better decisions than individuals or experts. Small groups risk making worse decisions because they may lack relevant information or a diverse range of viewpoints [19].

In an online context, large groups may be more manageable thanks to asynchronous communication tools that allow participation without requiring that everybody be present and paying attention. Intelligent user interfaces allow long discussions to be easily browsed or searched. However, the often impersonal nature of online interactions may serve to encourage uninhibited and antisocial behavior [11, 16], which may further reduce cooperation and increase conflict in large groups. This brings us to our first research question:

RQ1 ***Group Size.*** How does group size affect decision quality in online communities?

A crucial part of any decision-making process is defining the group that is responsible for making the decision. Ideally, a decision-making group should be a representative subset of the organization or community that needs the decision to be made. Traditionally, groups have been created in a top-down manner by an authoritative figure (i.e., a manager, or in academic studies, the experimenter). However, some groups, such as working groups or ad hoc committees, can be self-forming.

The manner in which self-formed groups attract and recruit participants can have a profound effect on group composition, which can in turn influence decision quality. For instance, group members may naturally choose to solicit those in their own social networks. Because social networks exhibit homophily and tend to be a source of behavioral homogeneity [14], the group composition may be skewed toward particular attitudes or preferences that are not representative of the community as a whole. Self-formed groups are increasingly common, especially in online communities [9], and in the current work, we look at one aspect of biased group recruitment in an online social production community, and explore how it affects decision quality.

RQ2 ***Group Formation.*** How does biased group formation affect decision quality in online communities?

The members of a decision-making group are likely to have differing levels of experience working with the organization or community. Some participants may be oldtimers with substantial experience, while others may be newcomers who are still learning about their roles. The group diversity literature suggests that such diversity can be both good and bad. The informational perspective hypothesizes that heterogeneous groups do better because they have a broader range of knowledge, skills, and opinions to draw from. Newcomers provide new ideas and perspectives, while oldtimers provide experience and structure. On the other hand, the social categorization perspective suggests that diversity is harmful because people use the differences to categorize group members into subgroups, which can lead to increased conflict and an adversarial "us versus them" dynamic [20].

In online communities, the effects of tenure diversity may be confounded by the fact that indicators of tenure and status are not always made salient. Many have hypothesized that masking these indicators reduces the effect of social categorization and status inequalities, which, in turn, can help equalize participation levels, promote communication openness, and improve decision quality. However, study results have been equivocal [1, 4]. Our next research question seeks to explore further the role of newcomer participation and tenure diversity on decision quality.

RQ3 ***Experience.*** How does newcomer participation and tenure diversity affect decision quality in online communities?

Finally, decision-making groups typically have an administrator or leader who is responsible for identifying and carrying out the group's decision. In some cases, that person may be partial to a particular outcome, and may use their influence to steer the group toward that outcome. Decisions made under such conditions are suspect because valid arguments and viewpoints may be ignored.

For instance, in [19], Surowiecki describes a crucial decision made by the NASA Mission Management Team during the space shuttle Columbia's final mission. The team met to decide whether to more thoroughly investigate the possibility that the shuttle sustained severe damage during launch. Surowiecki presents evidence that the team's leader had already made up her mind that the damage was inconsequential before the meeting, and deflected and downplayed issues brought up by engineers during the meeting. In essence, the leader ignored a number of valid concerns, leading to a flawed decision-making process and an incorrect decision that arguably resulted in the loss of the shuttle and its crew. Similar effects of leader bias have been reported in judicial decisions [15].

In online social production communities, biased administration may be even more of an issue as self-formed groups become more prevalent. Administrative roles in such groups are often volunteer-based or even self-appointed, which may be susceptible to selection biases, since volunteers may choose to seek power in areas where they have strong pre-formed opinions. Also, even if administrators endeavor to perform their duties in an impartial manner, they may still be affected by subconscious or hidden biases [8]. In the current work, we study the effects of apparent administrative bias on decision quality.

RQ4 ***Administrative Bias.*** How does biased group administration affect decision quality in online communities?

1.3 Decision Making in Wikipedia

Our overarching goal in asking these research questions is to understand how decision-making processes in online social production communities work and to learn how to improve their effectiveness through better processes, software tools, and intelligent interfaces. In the present work, we explore these questions in the context of one of the largest social production communities in the world: the English Wikipedia. With millions of contributors and articles, countless decisions must be made every day to keep the encyclopedia running smoothly. There has been substantial research in how Wikipedians successfully manage such a large community. Forte and Bruckman examined Wikipedia's self-governance, noting that there has been an increasing level of decentralization in its decision-making processes; decisions that were once reserved for founder Jimmy Wales are now made by the community [6]. Other research has examined more specific aspects of Wikipedia's decision-making processes, including how specialized tools enable vandal fighters to make decisions more efficiently [7], how Wikipedia's user promotion decisions compare to stated policy [3], and how the community decides which articles to feature on the front page [21].

One of the most important decisions that a social production community must make is to define the scope and breadth of the community's efforts. This issue is at the core of one of Wikipedia's long-running conflicts. Some believe that Wikipedia should be selective regarding what topics merit inclusion in the encyclopedia in order to ensure that it remains a maintainable and high-quality resource. Others disagree, arguing that Wikipedia ought to be inclusive and accept reasonable articles about anything that someone chooses to write about. They believe that doing so plays to

Wikipedia's strengths as a digital resource with no practical size limitations, and leads to more opportunities for people to contribute.

This conflict often manifests itself as disagreements over whether specific Wikipedia articles should be deleted. Lam and Riedl studied article deletions on Wikipedia at a broad level and found that as many as one-third of new articles are deleted [12]. In this paper, we address our four research questions by analyzing the processes that Wikipedians use to decide whether to keep or delete an article, and by looking at how different group composition factors influence decision quality.

2. DATA AND METHODS

2.1 Article Deletion on Wikipedia

In Wikipedia, deleting an article involves an extensive set of processes that is uncharacteristically *not* wiki-like. While anyone can create and edit an article, ordinary Wikipedians are limited to proposing and discussing deletions. Only administrators have the ability to delete an article. All article deletions must adhere to the procedures described in Wikipedia's "Guide to deletion," which leads users through choosing and invoking the appropriate deletion process.[1] Wikipedia's deletion processes as of early 2009 are summarized as follows.

Speedy deletion. This process is used for articles that are obviously inappropriate (e.g., vandalism or libel). If an article meets a speedy deletion criteria,[2] any user may mark it as such. Barring legitimate objection, an administrator will delete the article.

Proposed deletion. This process is used for uncontroversial deletions that do not meet the speedy deletion criteria. If a user proposes a deletion and nobody objects within five days, then an administrator will delete the article.

Articles for deletion (AfD). This process is used if the previous two processes do not apply or if there is objection to a speedy or proposed deletion. A user starts the AfD process by nominating an article to be deleted and providing a reason. Then, interested members of the community spend five or more days discussing the deletion. Finally, a neutral administrator examines the group discussion and determines what the community has decided to do. The administrator then takes the appropriate action, and closes the discussion. The typical outcomes are to delete or to keep the article.

We are interested specifically in the third process—articles for deletion, or AfD—because it involves the community coming together and making a collective decision about what to do with an article. We choose to study these particular decisions in Wikipedia because they are organized in a relatively standardized format that is amenable to automated coding, they occur frequently enough for our quantitative methods to be effective, and they involve self-formed groups with a sufficiently wide variety of compositions to address our research questions.

A typical AfD discussion and decision is shown in figure 1. Here, the user Merope has nominated the article *Lighthouses in Spain* for deletion because he believes that it is a trivial list of information that adds little value to Wikipedia. Over the next eight days, several others discuss whether the article should be deleted: Kwsn, C.Logan, and JForget favor deletion, while Dhaluza, Dhartung, Steve Hart, and Sjakkalle are opposed. Finally, administrator Akhilleus determines that the community has reached consensus to keep the article. He announces the result, thereby closing the discussion.

To the casual observer, it may appear as though the closing administrator may be merely tallying up how many participants "voted" for each outcome because the participants structure their arguments in a vote-like format. Each argument is prefixed with a clear and brief summary that is visually distinct and easily counted (e.g., "**Keep.**" or "**Delete.**"). However, the norm in Wikipedia is that AfD, along with most other decision-making processes, are *not* vote-based. Instead, Wikipedians expect administrators to carefully study the arguments and determine whether the participants have reached a "rough consensus" in deciding what to do.[3]

Wikipedians refer to these vote-like statements as "!votes" (read as "not-votes") as a tongue-in-cheek reminder that while the discussions may resemble votes, voting is not actually taking place, and opinions that are not accompanied by valid reasoning may be ignored. For brevity, we adopt similar nomenclature in this paper, referring to discussion participants as ***!voters*** and their preferred outcomes as ***!votes***. Also, when the meaning is clear from context, we use the term "AfD" to refer to specific instantiations of the Articles for Deletion process, rather than the process itself.

2.2 Measuring Decision Quality

A key part of exploring our research questions is knowing whether the AfD decisions being made are good. The usual scientific approach here might be to identify what factors Wikipedians use to evaluate whether an article belongs in Wikipedia, and to operationalize them as metrics that estimate these factors. For instance, in [12], Lam and Riedl found that the plurality of article deletions occur because the articles are about topics that do not sufficiently meet Wikipedia's notability guidelines. In their work, they defined metrics to estimate the notability of an article topic. In principle, we could use these metrics to help classify decisions as good or bad by looking for keep decisions made on low-notability articles, or delete decisions made on high-notability articles. However, we believe that such an approach is awkward for two reasons.

First, metrics of this nature are imprecise. The notability metrics in [12] are noisy and are difficult to apply definitively in individual cases. Wikipedians are aware of these metrics and do not consider arguments based on them to be valid. We believe that most extrinsic metrics of this type will be similarly unsuitable for rendering judgment on individual decision correctness.

Second, even if we had precise and accurate metrics, we are still left with the problem of defining a value as the threshold that an article must meet to avoid deletion. This is problematic because there is no gold standard. Part of Wikipedia's ethos is to allow its community to make its own decisions about content, style, and governance. There is no wrong decision as long as it was made in good faith as a way to move forward with the overarching goal of producing a free, high quality encyclopedia. It is difficult for us as outsiders to justify declaring that some AfD decisions were "wrong" just because a metric that we invented said so.

Instead, we measure decision quality by observing feedback in the system itself and looking for evidence that the community believed that a decision it previously made was incorrect. In the context of Wikipedia AfDs, we look for decisions that are reversed; that is, we find cases where an article is:

- deleted via AfD, but is re-created at a later date, or
- kept via AfD, but is deleted at a later date.

These reversals can occur through a variety of mechanisms. For instance, the decision may have been reversed due to a formal appeal lodged at one of Wikipedia's dispute resolution channels, or

[1] http://en.wikipedia.org/wiki/WP:GD

[2] http://en.wikipedia.org/wiki/WP:CSD

[3] http://en.wikipedia.org/wiki/WP:NOTAVOTE


The result was **keep.** --Akhilleus (talk) 20:43, 28 June 2007 (UTC)

Lighthouses in Spain [edit]

Lighthouses in Spain (edit | talk | history | links | watch | logs) – (View log)

zomg listcruft!!!111! Erm. Sorry. Wikipedia is not an indiscriminate list of information. -- Merope 17:42, 20 June 2007 (UTC)

- **Delete** Category is there already, no need for a list. Kwsn(Ni!) 17:51, 20 June 2007 (UTC)
- **Delete** - Per Kwsn. Why do we need two pages to do the work of one?--C.Logan 18:00, 20 June 2007 (UTC)
- **Delete** since the category exist, lots of red links too. But there at list 15-20 other list similar to that, so I guess most of them are listcruft as well.--JForget 19:34, 20 June 2007 (UTC)
- **Keep** This is Cruftcruft. First, the category only lists articles created--many of the lighthouses on the list will never have stand-alone articles, and we usually merge these to a list. Also lighthouses are prominent geographic landmarks, important in both marine navigation, and human culture. Each one must be unique for identification, and the unique characteristics are often associated with the adjacent settlements. Dhaluza 00:56, 21 June 2007 (UTC)
- **Keep**, a perfect example of a list that does what a category cannot, show articles that are not yet created. --Dhartung | Talk 04:38, 21 June 2007 (UTC)
  - **Comment** I'll give you that one, but how can the lighthouses be verified? That's a big concern of mine. Kwsn(Ni!) 17:21, 22 June 2007 (UTC)

    Look on a map. We don't delete articles because you can't verify it without getting out of your chair. Dhaluza 12:01, 28 June 2007 (UTC)
- **Note**: This debate has been included in the list of Spain-related deletions. -- John Vandenberg 09:16, 21 June 2007 (UTC)
- **Weak keep**, since we're using lists on WP, this one serves its purpose. -- Steve Hart 14:55, 25 June 2007 (UTC)
- **Keep**. Lighthouses are important features in ocean navigation, and therefore important geographical landmarks, often receiving a pretty prominent mark on maps. Sjakkalle (Check!) 10:48, 28 June 2007 (UTC)

Figure 1: A typical Wikipedia Articles for Deletion (AfD) discussion (from http://en.wikipedia.org/w/index.php?oldid=141246516).

an informal conversation among the involved participants. Alternately, a bold user or administrator might have simply taken the initiative to reverse a decision that he or she felt was incorrect. Since decision reversals can themselves be reversed upon community scrutiny, we only consider reversals that "stick"; that is, reversals that are persistent and are not undone. To help avoid cases where decision reversals may be due to policy changes or other long-term changes in the ecosystem, we only consider cases where an AfD decision is reversed within one year as being an indicator of a flawed decision. In addition, we do not consider cases where articles are re-created as a redirect (a pointer to another article) to be a flawed deletion decision.

We acknowledge that this is an imperfect method to measure decision quality. Not all bad decisions will be fixed by the community, and not all reversals are the result of flawed decisions. However, we feel that this approach represents an effective microscope into which decisions are of questionable quality, and allows us to study them without requiring us to impose our own judgment about the community's decisions.

2.3 Data Sources

To collect the requisite information to explore our research questions, we used the Wikimedia Foundation's data dumps[4] and the Wikimedia Toolserver.[5]

Current Versions Dump. We used the current versions dump, which contains the current text of every Wikipedia page, to obtain the text of all archived AfDs. Using the *mwlib* library to parse the wiki markup, we wrote a program that extracted the key information from each AfD: the article being discussed, the nominator's name, the participants' names and !votes, the closing administrator's name, and the decision. We discarded any discussions that did not appear to be in the de-facto standard format shown in figure 1.

To check the program's correctness, two people examined 48 random AfDs and noted any errors in the extracted data. The judges found errors in 3.8% of the pieces of collected information. Many errors involved the program misidentifying a participant's !vote, and were due to people expressing their !vote in unconventional ways, or making complicated arguments that could not be classified easily. We felt that an accuracy rate of over 95% was acceptable for a simple parser, and did not believe that more complex techniques such as sentiment analysis would be worth the added cost.

Metadata Dump and Event Log. The remainder of our data came from two sources: the historical revision metadata dump and the event log dump. The revision metadata dump tells us when each Wikipedia edit occurred, and who made each edit. The event log tells us when pages were deleted, restored, or renamed. We used these data sources for three purposes.

First, the data helped us verify and refine our automated coding program's outputs. Each AfD event (nomination, !vote, or closure) should correspond to an edit to the AfD discussion page at the time indicated in the user's signature. Using the revision metadata, we checked whether this was true for our collected data. If not, we checked whether a simple username or time correction would make the event consistent with the metadata (some Wikipedians' signatures contain alternate versions of their username or a non-UTC time). If that failed, we omitted the suspect event from our analysis. Additionally, we used the data to verify whether our program correctly assessed each AfD's result. For instance, if an AfD resulted in a delete decision, then a corresponding deletion should appear in the event log immediately following the AfD's closure.

Next, the data allowed us to detect whether an AfD decision was reversed, which, as described in section 2.2, is our indicator of an incorrect decision. For example, if an AfD resulted in a keep decision, we looked for evidence of a reversal by searching the event log for a deletion occurring after the discussion was closed. If we found one, we would then also search the data for the article's subsequent re-creation to determine whether the reversal was sticky.

Finally, we used the metadata to compute other metrics about each AfD that we use in our model of decision quality. Since the metadata dump omits deleted articles, we used the Wikimedia Toolserver as needed to obtain metadata about deleted articles. We will describe these metrics in sections 2.5 and 3.

2.4 AfD Data Set

Our data set contains 158,733 AfDs occurring between January 1, 2005 and April 1, 2009. We chose to exclude discussions starting before January 1, 2005 because many of them were not in the

[4]http://en.wikipedia.org/wiki/WP:DUMP
[5]http://meta.wikimedia.org/wiki/TS

format depicted in figure 1, and thus could not be processed by our automated coding tool.

Most AfD decision-making groups are small—the median number of !voters (not including the nominator) is four. Because we are interested in studying *group* decision making, we omit 12,997 AfDs that had zero or one !vote from our analysis.

A majority of AfDs, 68%, result in a decision to delete the article, while 25% result in the article being kept. The remaining 7% represents a variety of uncommon outcomes, including merging the article's content into another article, redirecting readers to another article, or moving the content to another wiki. Because it is unclear how to identify whether such decisions are reversed, we discard this 7%, which leaves a total of 135,461 AfDs in our analysis. Of these, we found that 4.67% of delete decisions and 3.52% of keep decisions are reversed.

At first glance, the fact that over two-thirds of discussions result in deletion may make AfD look like an unfairly biased process. However, the disparity is perhaps expected: it is reasonable to believe that someone would only nominate an article for deletion if there were good reasons for doing so (thus, making deletion a likely outcome). Someone who nominates articles haphazardly might face consequences for being disruptive.

2.5 Modeling Decision Quality

Our analysis of decision quality uses a logistic regression model. The binary dependent variable is whether the AfD decision is reversed, and the independent variables represent properties of each decision-making group. To account for factors that may correlate with reversals but that are not related to the factors that we are studying, we control for several constructs as described below:

Temporal effects. As Wikipedia and its users age, there may be natural changes in how often decisions are reversed resulting from factors such as policy changes or broad shifts in community behavior. We control for this by including a *DiscussionDate* variable, the date that the AfD discussion was started, in the model, expressed as the number of years after January 1, 2005. Additionally, we control for any effects due to the newness or staleness of the nominated article, given as *ArticleAge*, the article's age in days at the time it was nominated for deletion.

Strength of consensus. There is reason to believe that decisions made through weak consensus are more likely to be reversed than those with strong or unanimous consensus. To model the strength of consensus, we use the percentage of participants who !voted for the eventual decision (*ConsensusStrength*). A value of one indicates a "unanimous" consensus, while lower values indicate weaker consensus and (perhaps) increased controversy.

Stakeholder impact. The stature and size of the groups affected by a decision and their participation in making the decision may have an impact on whether the decision is reversed. In the context of AfDs, the user(s) who authored the nominated article likely feel the most affected by the debate since the community is considering throwing their work away. We use three variables to control for this: the number of users who contributed to the article before its nomination (*NumEditors*), the experience of the user who created the article (*CreatorEdits*, defined as how many Wikipedia edits the user had made before creating the article), and whether he or she was involved in the AfD discussion (*CreatorVoted*).

Decision outcome. Different decisions may be more or less likely to be reversed depending on factors such as group attitudes, norms, and procedures. Here, reversing a keep decision may require the community to go through the deletion processes again, while reversing a delete decision may entail rewriting the article. These differ in difficulty, formality, and level of community scrutiny.

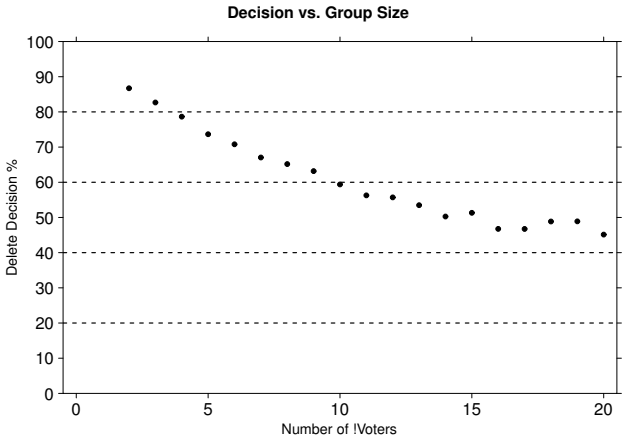

Figure 2: Relationship between AfD decision and group size.

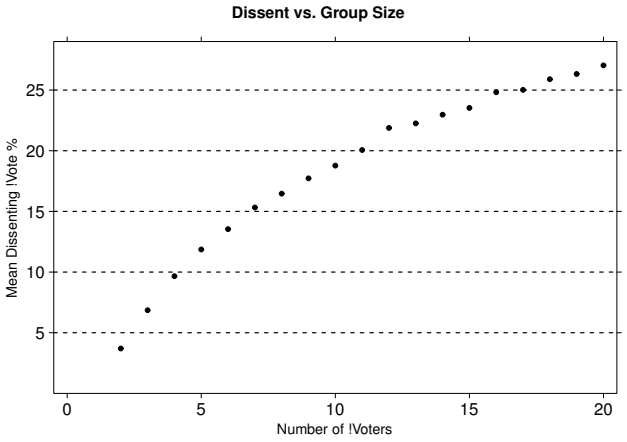

Figure 3: Relationship between group dissent and size in AfDs.

Since these are fundamentally different processes, we believe that our research questions may have answers that depend on which decision was made. To this end, we present our results as two separate models: one for AfDs that resulted in a delete decision, and one for AfDs that resulted in a keep decision.

3. ANALYSIS AND RESULTS

Now, using the AfD data set and our models of decision quality, we explore each of our four research questions. In each subsection, we will present analysis of policy or exploratory data to motivate interesting hypotheses, state our hypotheses, explain how we tested each hypothesis, and describe the results from our full models.

Table 1 shows our two models of decision quality, along with descriptive statistics about the input variables. Control variables are described in section 2.5, and independent variables are described in the following subsections. All variables have a variance inflation factor (VIF) of below 2.5, which suggests that inflated standard errors due to multicollinearity is not an issue [10]. Since the distribution of several variables is right-skewed, often with standard deviations larger than the mean, we apply base-2 logarithms to transform them to approximately-normal variables. These variables are labelled with "log2" in table 1.

3.1 RQ1: Group Size

We begin by exploring the question of whether, as suggested by research on offline group dynamics, group size affects the AfD decision process. Figure 2 indicates that different-sized groups tend to yield different decisions. Large groups make fewer decisions to

Model 1: Delete decisions					Model 2: Keep decisions			
Mean	**S.D.**	β	**Odds Ratio**	**Variable**	**Mean**	**S.D.**	β	**Odds Ratio**
–	–	-1.8233	0.161 ***	Intercept	–	–	-0.6152	0.541 ***
2.08	1.09	-0.0303	0.970 +	DiscussionDate	2.32	1.11	-0.4649	0.628 ***
156	285	0.0035	1.004	ArticleAge (log2)	346	439	-0.0447	0.956 ***
576	3861	0.0352	1.036 ***	CreatorEdits (log2)	2094	8163	-0.0095	0.990
10.67	40.34	0.1409	1.151 ***	NumEditors (log2)	35.6	142	0.2401	1.271 ***
0.0556	0.229	-0.2065	0.813 **	CreatorVoted (0/1)	0.159	0.366	0.4086	1.505 ***
0.909	0.161	-1.9025	0.149 ***	ConsensusStrength (%)	0.801	0.199	-3.4552	0.032 ***
5.43	4.20	-0.1159	0.891 ***	H1: GroupSize (log2)	8.22	7.40	-0.1597	0.852 ***
47.1	187	0.0227	1.023 *	H1: GroupSizeSq	122	725	0.0365	1.037 *
0.00396	0.0628	0.0401	1.041	H2: BotRecruit (0/1)	0.0137	0.116	0.1385	1.149
0.0171	0.130	0.1121	1.119	H2: NomRecruit (0/1)	0.0374	0.190	-0.2213	0.802
0.00795	0.0888	-0.0850	0.918	H2: DeleteRecruit (0/1)	0.00635	0.0794	0.0562	1.058
0.00280	0.0528	0.3966	1.487 *	H2: KeepRecruit (0/1)	0.0357	0.185	-0.2448	0.783
0.0756	0.134	0.0894	1.094	H3a: AfDNewcomers (%)	0.119	0.151	0.5278	1.695 **
0.0203	0.0711	0.4108	1.508 *	H3a: WPNewcomers (%)	0.0262	0.0754	2.3611	10.603 ***
0.591	0.270	0.0003	1.000	H3b: TenureDiversity	0.608	0.254	-0.0945	0.910 **
0.422	0.342	0.0003	1.000	H3b: TenureDiversitySq	0.434	0.324	-0.0091	0.991
0.0544	0.227	0.0814	1.085	H4: AdmDeleteBias (0/1)	0.0260	0.159	-0.0100	0.990
0.193	0.394	-0.0826	0.921 +	H4: AdmKeepBias (0/1)	0.127	0.333	0.2051	1.228 **
			1132.29 ***	Likelihood Ratio				1142.14 ***

Table 1: Descriptive statistics of variables and results of logistic regression predicting flawed decisions. *Negative* β values and odds ratios *below* 1 indicate variables associated with better decision quality. (* $p < .001$, ** $p < .01$, * $p < .05$, + $p < .1$)**

delete an article than average, while small groups make more decisions to delete an article than average. Figure 3 shows that as group size grows, so does the average percentage of people who dissented and !voted against the eventual decision. These relationships suggest that group size has a fundamental effect on how decisions are made, and lead us to believe that it may have an effect on decision quality. We believe larger groups will benefit from additional viewpoints and information, but with diminishing returns since conflict and dissent will also become increasingly prevalent.

H1 *Bigger-Better*: Larger groups will make better decisions than small groups, but with diminishing returns.

We measured the effect of group size on decision quality by introducing *GroupSize*, a normalized variable containing the number of !voters in the AfD. To test for non-linear effects, we also added the quadratic term *GroupSizeSq*.

Model 1 and 2 both show that group size and its quadratic term have significant effects on whether a decision is reversed. Figure 4 depicts the effects from both models, which are similar to one another and show that delete decisions made by small groups are more likely to be reversed than those made by larger groups. The plots flatten out toward the right, suggesting that there is little benefit from increases in size once a group is moderately sized.

3.2 RQ2: Group Formation

Wikipedia's AfD decision-making groups are self-formed, which, as described earlier, carries a risk of biased recruitment. Also, group members could, in principle, strategically choose who to recruit in an attempt to influence the decision-making process. This is particularly true when group sizes are small and the addition of one or two people could sway consensus, which is the case with Wikipedia AfD discussions (recall that the median !voter count is four). To help avoid this sort of antisocial behavior, Wikipedia's norms and policies only allow a limited form of direct recruitment. For AfDs, they permit neutral recruitment of members of two groups: the nominated article's primary contributors, and relevant WikiProjects (work groups that focus on particular topics).[6]

[6]http://en.wikipedia.org/wiki/WP:AFDHOWTO

However, we observe that this policy itself has a form of bias. The permitted groups are comprised of people who have an interest in the nominated article, and thus may be predisposed to resist efforts to delete the article. There is valid reasoning for the policy though: because these groups' members are likely those who know the most about the article's topic, they are the most able to help make a well-informed decision, and may be able to address any deficiencies in the article that caused its nomination for deletion.

That said, it remains the case that Wikipedia's recruitment policy contains bias. Also, the policy may not be strictly enforced. There is no automated means of detecting improper recruitment, and manually investigating every participant is labor-intensive and draconian. We hypothesize the following about AfD decision-making groups that are formed through recruitment:

H2 *Recruit-Worse*: Groups formed through recruitment make worse decisions than naturally formed groups.

To test this hypotheses, we first need to observe and measure recruitment to AfD discussions. On Wikipedia, the typical method to communicate with a user is with *User Talk* pages, which are wiki pages associated with user accounts. So, to recruit a user to join a discussion, one would edit that user's *User Talk* page and write a recruitment message. We detect cases of successful recruitment to AfD discussions by processing the metadata dump described earlier, looking for instances of the following sequence of events.

1. User *A* participates in AfD discussion *D*, either by nominating an article for deletion (thereby starting the discussion), or by expressing a !vote.
2. Within one hour or ten edits (whichever sooner) of (1), user *A* edits user *B*'s User Talk page.
3. Within two days of (2), user *B* !votes in discussion *D*.

When we find such a sequence of events, we say that user A has successfully recruited user *B* to participate in AfD discussion *D*. We note that this is not definitive evidence of recruitment since it is possible that *A*'s message to *B* is unrelated to the AfD discussion, and that this sequence occurred due to coincidence. However, we believe this approach works reasonably well in practice, and is easily automated. Furthermore, we also apply the following heuristics

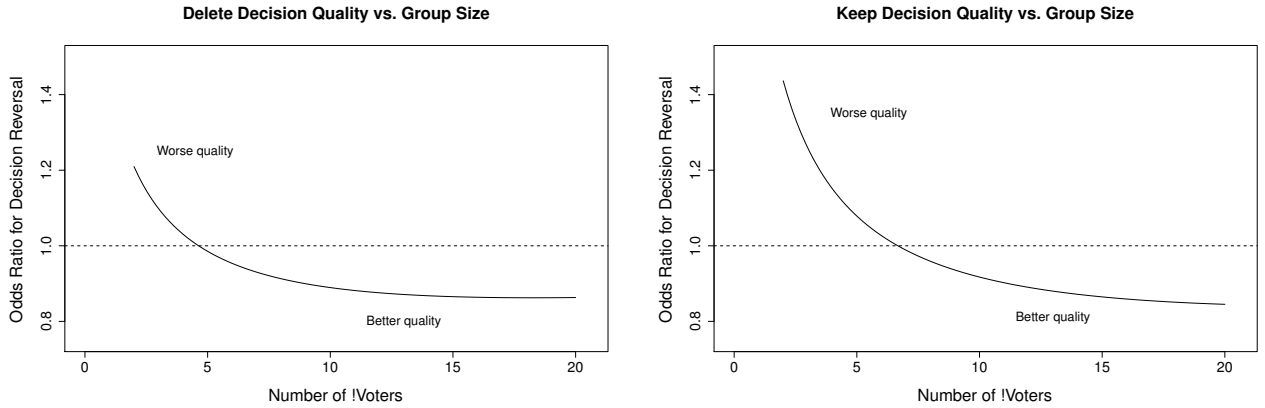

Figure 4: Effect of Group Size on Decision Quality (H1). Lower odds ratios indicate better decision quality.

Recruiter	AfD Nom.	Delete !Voter	Keep !Voter	Bot
% of Recruits who !Voted Delete	34.6%	60.7%	15.1%	20.3%

Table 2: Summary of AfD group recruitment showing how recruited participants' !votes differ depending on who recruited them. Wikipedia-wide, 62% of AfD !votes are for deletion.

to help filter out common false positives that we observed while performing manual spot-checks of the data.

First, if *A* or *B* had edited each others' User Talk page in the three days before the supposed recruitment message, then we did not consider the instance to be recruitment because *A* and *B* were probably having an unrelated conversation.

Second, if *B* was active in more than three other AfD discussions in the three days before or one hour after his or her !vote in *D*, then we did not consider the instance to be recruitment. We observed that active AfD participants tend to also be active elsewhere in Wikipedia. They receive many messages on their User Talk pages, including ones from peers who participated in a similar set of AfDs. However, the messages often are not AfD recruitment messages, but are thank yous, warnings, feedback, or commentary regarding various topics that the users were involved with.

We also found that there have been two bots (computer programs that edit Wikipedia)—*BJBot* and *Jayden54Bot*—that automatically automatically notified article editors about AfD discussions and recruited them to participate per the established policy. These bots performed AfD notifications for several months, and offer us an opportunity to study the effect of recruitment that is purely policy driven. We use a process like one described above to detect successful instances of bot-initiated recruitment: if a recruitment bot edited a user's talk page, and that user !voted in an AfD within two days, then we consider that user to have been recruited by the bot.

Using the above processes, we identified 8,464 instances of successful recruiting. Table 2 shows a summary of who did the recruiting, and how their recruits !voted. We see large differences in !voting behavior, which suggests that there is bias in who people choose to recruit. (From these data we cannot tell whether the bias is an intentional effort to influence consensus, or the result of social network homophily [14].) Participants recruited by keep !voters were about four times less likely to support deletion as those recruited by delete !voters. The participants that bots recruited also appear unlikely to support deletion, which reflects the policy bias we observed earlier.

To see what effect participant recruitment has on decision quality, we introduce four binary variables: *BotRecruit*, *NomRecruit*, *DeleteRecruit*, and *KeepRecruit*. These variables indicate whether a bot, the AfD nominator, a delete !voter, or a keep !voter successfully recruited somebody to the group, respectively.

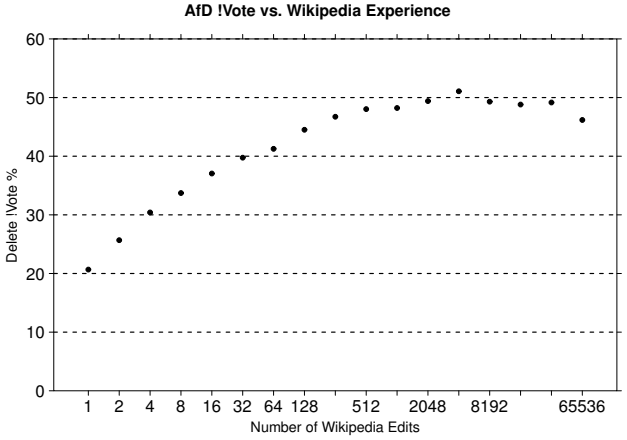

Figure 5: Relationship between AfD !vote and Wikipedia experience at time of !vote, computed on a per-user basis.

Looking back to table 1, we find that regardless of the decision, none of the first three variables has a statistically significant effect. On the other hand, when a keep !voter recruited someone to the discussion, we see a significant effect: delete decisions are more likely to be reversed. We offer two possible explanations: the first is that recruitment by keep !voters, biased as it may appear, is a sign of positive community interest, and suggests that the article should be kept. If the community decides otherwise and deletes the article, then decision quality suffers. An alternative explanation is that keep !voter recruitment is a sign of activism among those who prefer to keep the article. These proponents may be especially persistent in maintaining the article's existence in Wikipedia, even if it requires working to reverse a delete decision.

3.3 RQ3: Experience

Next, we turn to our research question regarding the role of participant experience and tenure diversity in decision quality. Our exploration of Wikipedia AfD discussions suggests that there are fundamental differences between newcomer and oldtimer behavior. As shown in figure 5, newcomers are far less likely than experienced users to support deleting a nominated article. Newcomers' knowledge and interpretation of Wikipedia's article policies are evidently different from that of oldtimers, perhaps due to lapses in newcomer

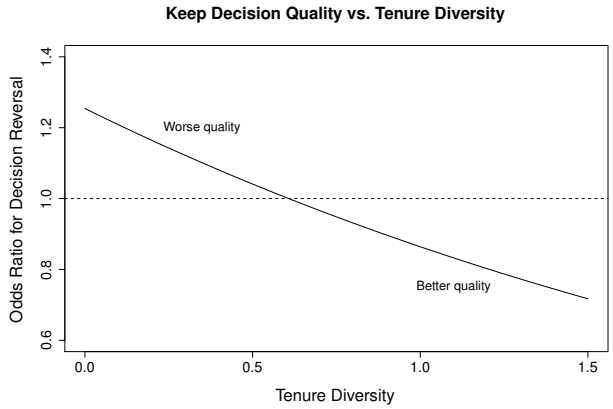

Figure 6: Effect of Tenure Diversity on Decision Quality (H3b). Lower odds ratios indicate better decision quality.

socialization. This suggests that newcomer participation in AfD discussions may adversely affect decision quality.

By contrast, diversity theory says that increased tenure diversity can lead to better group outcomes. A recent study by Chen, et al. showed that moderate tenure diversity in Wikipedia groups is beneficial to productivity and retention [4], and we believe the effect will extend to decision quality. We hypothesize:

H3a *Newcomers-Worse*: Groups with more newcomers make worse decisions than groups with fewer newcomers

H3b *Diversity-Moderate*: Groups with moderate tenure diversity make better decisions than groups with high or low tenure diversity

To test H3a, we introduce two measures of newcomer participation to our model: percentage of participants who had 15 or fewer Wikipedia edits (*WPNewcomers*), and percentage of participants who were not new to Wikipedia, but had !voted in five or fewer AfDs (*AfDNewcomers*). To test H3b, we include the normalized tenure diversity of the participants who !voted in the AfD discussion (*TenureDiversity*), and its quadratic term (*TenureDiversitySq*). Our definition of tenure diversity is identical to the one used by in [4]: the coefficient of variation of the number of days since each user's first Wikipedia edit. The coefficient of variation is a widely used measure of tenure diversity in past research [2].

We start by looking at H3a. Both models in table 1 show that decisions made by groups that have Wikipedia newcomers are significantly more likely to be reversed. Model 2 shows that participation by AfD newcomers also leads to more reversals when they are involved in making a keep decision. Proceeding to H3b, we turn to the tenure diversity measures. We find that changes in tenure diversity have no effect when the decision is to delete. However, diversity has a significant effect when the decision is to keep. A plot of the effect on odds ratio is shown in figure 6, and indicates that decision quality improves with tenure diversity. It is unclear why tenure diversity's effect appears dependent on the decision outcome.

3.4 RQ4: Administrative Bias

Finally, we look at RQ4, which is about the effect of administrative bias on decision quality. In Wikipedia AfDs, the administrator who closes the discussion is responsible for identifying what the community has decided to do. To do this, the administrator applies guidelines that describe how to interpret the discussion and determine whether a rough consensus was reached.[7]

The guidelines allow for some subjectivity. They require that the administrator to use his or her "best judgment," but to "be as impartial as is possible for a fallible human." Because hundreds of administrators volunteer to close AfD discussions, and because determining consensus requires human judgment, there exists opportunity for administrative bias to affect the AfD decision-making process. We expect that groups with biased administrators will yield poorer decision quality than those with unbiased ones.

H4 *Biased-Admin-Worse*: Groups with biased administration make worse decisions than groups with neutral administration

We measure administrative bias by looking at how different administrators make consensus calls in AfDs that have similar !vote breakdowns, and comparing their behavior to a Wikipedia-wide statistic. To provide some intuition, let us consider AfDs in which three participants !voted keep and four !voted delete. Historically, 56% of such AfDs have resulted in a decision to delete the nominated article. Now, suppose that Fred, an administrator, has closed ten of these AfDs, and that he determined there was a consensus to delete the article in two of them, or 20%. Since 20% is much less than 56%, the evidence suggests that Fred is biased away from delete outcomes and towards keep outcomes. (Administrators may also be biased toward certain topics or articles, but we do not consider such biases in the current work.)

To turn this approach into a bias metric, we first divide all AfDs into 11 groups based on each AfD's !vote breakdown, measured as percentage of !votes in favor of deletion. We ignore !votes that are for outcomes other than keep and delete (96% of !votes are to keep or delete the article). The first group contains AfDs with fewer than 5% delete !votes. The second group contains AfDs with at least 5% but fewer than 15% delete !votes. The third group contains AfDs with at least 15% but fewer than 25% delete !votes, and so on. The final group contains AfDs with at least 95% delete !votes.

Now, we can define a relative bias measure for an administrator A by summing the differences between his or her consensus calls and the Wikipedia-wide ones for each group. Because administrators may have little or no data in some groups of AfDs, we apply a form of Bayesian smoothing that is based on Wikipedia-wide statistics.

$$bias_A = \sum_{i=1}^{11} \left(\frac{C * p_i + numdel_{A,i}}{C + numafds_{A,i}} - p_i \right) \quad (1)$$

Here, p_i is the Wikipedia-wide percentage of AfDs in group i that were closed with a delete decision, $numafds_{A,i}$ is the number of AfDs in group i that A closed, and $numdel_{A,i}$ is the number that A closed with a delete decision. C is a tunable parameter used for smoothing. For our analysis, we set $C = 3$. Using this definition, we compute a bias for all administrators who have closed at least ten AfD discussions. Such administrators collectively closed 91% of the AfDs in our data set.

If the bias measure is positive, then the administrator tends to close discussions with a delete decision more often than average, given some !vote breakdown. Similarly, if the bias measure is negative, he or she is more likely to close discussions with a keep decision. The measure's magnitude represents how strong the apparent bias is. To make this measure easier to interpret, we normalize these values to standard scores. For example, a measure of -1.4 indicates that the administrator's apparent bias is 1.4 standard deviations from the average, and that the bias is in the keep direction.

To illustrate the differences that we see among administrators, figure 7 shows the consensus calls profiles for two administrators who appear diametrically biased with bias measures of -2.92 and +2.58. For comparison, we have included the Wikipedia-wide average consensus call profile. These plots show the likelihood that an AfD results in a delete decision for each of the 11 groups. Indeed, there appears to be substantial variation in how different administrators determine consensus.

[7] http://en.wikipedia.org/wiki/Wikipedia:DGFA

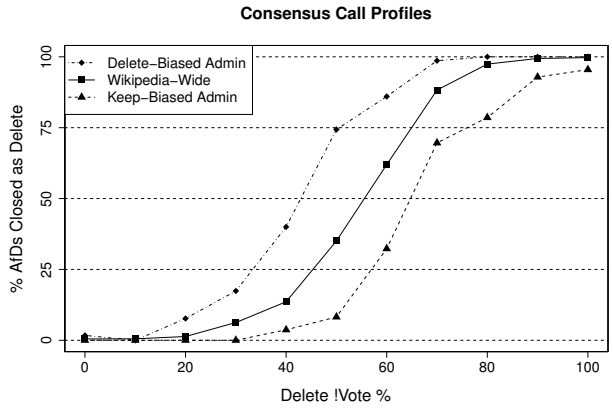

Figure 7: Relationship between !vote breakdown and decision.

Recall that Wikipedia's policy regarding decision making is that consensus should *not* be determined according to vote counts. However, we observe that the plots in figure 7 resemble logistic curves, suggesting that Wikipedia's rough consensus process approximates a vote, but permits administrators to, at their discretion, accept a compelling minority opinion as the decision. It is interesting to note that on average, there is a slight tilt toward keep decisions; the solid line in figure 7 shows that it requires more than 50% of !voters in an AfD to favor deletion before the likelihood of a deletion decision reaches 50%.

These measurable aggregate-level differences among administrators in when they accept minority opinions are suggestive of a bias in which some administrators are systematically discounting certain opinions, perhaps subconsciously. To determine the effect of this apparent bias in AfD discussions, we introduce a categorical variable that denotes whether the closing administrator is keep-biased (bias of less than -2), delete-biased (bias of greater than +2), or neutral (bias of -2 to +2, or uncomputable). We encode this as two dummy-coded variables, *AdmKeepBias* and *AdmDeleteBias*, with neutral administrators as the reference group.

Model 1 indicates that when keep-biased administrators are involved in delete decisions, there is a marginally significant decrease in reversals ($p = 0.0594$). If an administrator makes a consensus call that is contrary to his or her own bias, the community's arguments were likely strong enough to overcome that bias, thus leading to an increase in decision quality. On the other hand, model 2 shows that when these keep-biased administrators make the call to keep an article, their decisions are reversed more often. We see no significant effect on decision quality when delete-biased administrators are involved in decision making, but the models show weak trends consistent with our results for keep-biased administrators.

4. DISCUSSION AND CONCLUSION

In each part of section 3, we looked at the results of our analysis as they relate to one of our research questions. Now, we will step through the research questions and discuss our findings, as well as the implications for social production community design.

RQ1: Group Size. We find support for **H1** ***Bigger-Better***. Online decision-making processes that involve too few people are at higher risk of making low quality decisions. Larger groups make better decisions, but with rapidly diminishing returns.

Our findings lead us to two design suggestions. First, encourage more users to participate in collaborative decision-making activities. The increased group sizes can improve decision quality. However, since we see evidence of diminishing returns, it may be beneficial to steer users toward underpopulated areas instead of toward areas that are already crowded. Second, be wary of decisions that are made by groups that are very small, as they may be suspect. Scrutinize the decisions carefully, and consider delaying the decision to find additional participants if there does not appear to be a sufficient quorum. In communities where decisions are made in a structured manner, it may be possible to automate both suggestions through intelligent task routing techniques [5].

RQ2: Group Formation. In our exploratory analysis, we found strong evidence of biased recruitment to AfDs. People appeared to seek out like-minded peers. Despite the biases, our results only show limited support for **H2** ***Recruit-Worse***—decision quality is unaffected by most forms of recruitment that we studied. More work is needed to understand why this is the case. Perhaps Wikipedians are aware of these biases and are able to adjust accordingly.

However, our results do shed some light on the complexity that exists when decision-making groups are self-formed. Designers should carefully consider possible biases when constructing policy about how to attract participants (e.g., Wikipedia's policy bias as described in section 3.2). To reduce the amount of human effort required for recruitment, and to help avoid biases from selective recruitment, communities may wish to consider automating basic outreach strategies such as Wikipedia's AfD notification bots. Additionally, it may be possible to construct automated tools that look for signs of biased group formation, and to carefully scrutinize decisions made by such groups.

RQ3: Experience. We find partial support for **H3a** ***Newcomers-Worse*** and **H3b** ***Diversity-Moderate***. Newcomer participation is detrimental to decision quality, while high tenure diversity is beneficial in some cases. The latter finding partially disagrees with what diversity theory predicts, and suggests AfD groups do not suffer from the negative social categorization effects of high tenure diversity [4, 20]. A possible explanation for this is that AfD groups are ephemeral and highly task-oriented. Thus, conflict between newcomers and oldtimers might not appear in the AfD discussion, but may negatively affect their interactions elsewhere on Wikipedia.

In light of these findings, we recommend that social production communities should encourage all users (including newcomers) to participate in group decision-making processes, but they should focus on socializing newcomers to help them understand issues related to the decisions at hand. Automated tools could also watch for and draw attention to situations that lack sufficient newcomer or oldtimer presence. Taking steps to increase group diversity in these cases may provide improved decision quality as well as provide opportunities for newcomer socialization.

Looking back at figure 5, we are intrigued by the extent to which newcomers and oldtimers apparently differ in their opinions in AfDs. We speculate that in cases like this where newcomers tend to disagree with oldtimers, it could be beneficial to give serious thought to newcomer input and consider ways to integrate their ideas into community norms. Rebuking newcomers' opinions in favor of existing group ideals could alienate and drive away newcomers, which may be harmful in the long term. In Wikipedia's case, this may be a contributor to recent slowdowns in growth [18]. Established norms can certainly be difficult to change, but acceptance of new ideas may be necessary to keep a community sustainable.

RQ4: Administrative Bias. We find support for **H4** ***Biased-Admin-Worse***. The presence of biased administration can lead to worse quality when they are involved in decisions that agree with their bias. However, when they are involved in decisions contrary to their bias, we find evidence that decision quality improves. These findings lead us to two design recommendations.

First, builders and maintainers should be conscious of the pos-

Hypothesis	Result	Description
H1 ***Bigger-Better***	Supported	Larger groups make better decisions, but with diminishing returns
H2 ***Recruit-Worse***	Mixed	Biased recruitment leads to worse decisions under some circumstances
H3a ***Newcomers-Worse***	Supported	Newcomer participation yields worse decision quality
H3b ***Diversity-Moderate***	Mixed	Diverse groups may make better decisions; no social categorization effects were observed
H4 ***Biased-Admin-Worse***	Supported	Worse decisions in some cases if decision agrees with administrator's bias Better decisions in some cases if decision is contrary to administrator's bias

Table 3: A summary of our findings.

sibility of administrative bias. There should be a process for the community to challenge questionable decisions and inconsistent administration. For example, Wikipedia AfD decisions can be appealed through a process called Deletion Review. Such processes will enable greater community scrutiny of potentially problematic individuals, and may lead to improved decision quality.

Second, consider automated mechanisms that draw attention to contentious cases, especially if they involve an administrator who has acted in concordance with a history of apparent bias. By introducing additional community discussion or analysis by a secondary administrator, it may be possible to reduce the negative impact of any biases that might exist and effect better decisions.

Summary. In this paper, we have explored how four group composition factors influence decision quality in a large online social production community. Our findings are summarized in table 3. Earlier in this section, we provided discussion and recommendations that we hope will inform the design of more effective decision-making processes and tools. While not all of our findings are definitive, we believe they raise interesting questions for social production communities (e.g., how should a group fairly canvass the community for useful input, or address inconsistencies in administration?), and they point the way toward future work.

Our work focused on one class of content decisions made on Wikipedia, and thus, our results may not generalize to other types of decisions or other communities. Further work is necessary to test our results in different environments. Nonetheless, we believe that the decisions we studied—ones of content relevance and appropriateness to the community—are representative of decisions that communities typically face, and that our contributions here will help drive the development and evolution of future social production communities.

5. ACKNOWLEDGEMENTS

We are grateful for the support of the members of GroupLens Research. We thank the Wikimedia Foundation for their support of research through their release of data sets and assistance in making the Wikimedia Toolserver possible. This work is funded by the NSF, grants IIS 05-34420, 07-29344, and 08-08692.

6. REFERENCES

[1] B. B. Baltes, M. W. Dickson, M. P. Sherman, C. C. Bauer, and J. S. LaGanke. Computer-Mediated communication and group decision making: A Meta-Analysis. *Organ Behav Hum Dec*, 87(1):156–179, 2002.

[2] A. G. Bedeian and K. W. Mossholder. On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods*, 3(3):285–297, July 2000.

[3] M. Burke and R. Kraut. Mopping up: modeling Wikipedia promotion decisions. In *Proc. CSCW 2008*, pages 27–36, San Diego, CA, USA, 2008. ACM.

[4] J. Chen, Y. Ren, and J. Riedl. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proc. CHI 2010*, pages 821–830, Atlanta, GA, USA, 2010. ACM.

[5] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Proc. IUI 2007*, pages 32–41, Honolulu, HI, USA, 2007. ACM.

[6] A. Forte and A. Bruckman. Scaling consensus: Increasing decentralization in Wikipedia governance. In *Proc. HICSS 2008*, page 157. IEEE Computer Society, 2008.

[7] R. S. Geiger and D. Ribes. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proc. CSCW 2010*, pages 117–126, Savannah, GA, USA, 2010. ACM.

[8] A. G. Greenwald and M. R. Banaji. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995. PMID: 7878162.

[9] J. R. Hackman and N. Katz. Group behavior and performance. In *Handbook of Social Psychology*, volume 2, pages 1208–1251. Wiley, New York, 5th edition, 2010.

[10] R. R. Hocking. *Methods and applications of linear models*. John Wiley and Sons, Mar. 2003.

[11] S. Kiesler, J. Siegel, and T. W. McGuire. Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10):1123–1134, 1984.

[12] S. K. Lam and J. Riedl. Is Wikipedia growing a longer tail? In *Proc. GROUP 2009*, pages 105–114, Sanibel Island, FL, USA, 2009. ACM.

[13] J. M. Levine and R. L. Moreland. Progress in small group research. *Annu Rev Psychol*, 41(1):585–634, 1990.

[14] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annu Rev Sociol*, 27(1):415–444, 2001.

[15] E. A. Posner. Does political bias in the judiciary matter?: Implications of judicial bias studies for legal and constitutional reform. *U Chi L Rev*, 75(2):853–883, 2008.

[16] J. Siegel, V. Dubrovsky, S. Kiesler, and T. W. McGuire. Group processes in computer-mediated communication. *Organ Behav Hum Dec*, 37(2):157–187, 1986.

[17] I. Steiner. *Group process and productivity*. Academic Press, New York, 1972.

[18] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: Slowing growth of Wikipedia. In *Proc. WikiSym 2009*, pages 1–10, Orlando, FL, 2009. ACM.

[19] J. Surowiecki. *The Wisdom of Crowds*. Anchor, Aug. 2005.

[20] D. van Knippenberg, C. K. W. D. Dreu, and A. C. Homan. Work group diversity and group performance: An integrative model and research agenda. *The Journal of Applied Psychology*, 89(6):1008–1022, Dec. 2004. PMID: 15584838.

[21] F. Viegas, M. Wattenberg, and M. McKeon. The hidden order of Wikipedia. In *Proc. OCSC 2007*, pages 445–454. 2007.