Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression

Nathan TeBlunthuis
nathante@uw.edu
University of Washington
Seattle, Washington, USA

ABSTRACT

Organizing complex peer production projects and advancing scientific knowledge of open collaboration each depend on the ability to measure quality. Article quality ratings on English language Wikipedia have been widely used by both Wikipedia community members and academic researchers for purposes like tracking knowledge gaps and studying how political polarization shapes collaboration. Even so, measuring quality presents numerous methodological challenges. The most widely used systems use labels on discrete ordinal scales when assessing quality, but such labels can be inconvenient for statistics and machine learning. Prior work handles this by assuming that different levels of quality are "evenly spaced" from one another. This assumption runs counter to intuitions about the relative degrees of effort needed to raise Wikipedia encyclopedia articles to different quality levels. Furthermore, models from prior work are fit to datasets that oversample high-quality articles. This limits their accuracy for representative samples of articles or revisions. I describe a technique extending the Wikimedia foundations' ORES article quality model to address these limitations. My method uses weighted ordinal regression models to construct one-dimensional continuous measures of quality. While scores from my technique and from prior approaches are correlated, my approach improves accuracy for research datasets and provides evidence that the "evenly spaced" assumption is unfounded in practice on English Wikipedia. I conclude with recommendations for using quality scores in future research and include the full code, data, and models.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; *Social content sharing*; **Computer supported cooperative work**.

KEYWORDS

sociotechnical systems, measurement, statistics, quality, machine learning, peer production, Wikipedia, online communities, methods, datasets

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Nathan TeBlunthuis. 2022. Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Measuring content quality in peer production projects like Wikipedia is important so projects can learn about themselves and track progress towards goals. Measuring quality also helps build confidence that information is accurate and verifiable and supports monitoring how well an encyclopedia includes diverse subject areas to identify gaps where coverage of certain topics needs attention [31]. Once important gaps are identified, measuring quality enables tracking and evaluating the progress of subprojects and initiatives organized to fill them [16, 41]. Assessing quality also helps motivate contributors in that raising an article to a high standard of quality is a recognized achievement [5, 14]. In these ways, measuring quality can be of key importance to advancing the priorities of the Wikimedia movement and is also important to other kinds of open collaboration like open source software [10].

Measuring quality also presents methodological and ontological challenges. How can "quality" be conceptualized to allow precise and accurate measurement of the goals of a project and the value it produces? Language editions of Wikipedia, including English, have systems for quality assessment that have been useful both for motivating and coordinating project work and for enabling research. Epistemic virtues of this approach stem from the community-constructed criteria for assessment and from formalized procedures for third-party evaluation organized by Wikiprojects. These systems also have two important limitations: (1) ratings are likely to lag behind changes in article quality, and (2) quality is assessed on a discrete ordinal scale, which raises issues for statistical and machine learning modeling. Both limitations are surmountable.

The machine learning framework introduced by Warncke-Wang et al. [42], further developed by Halfaker [16], implemented by the Objective Revision Evaluation Service[1] (ORES) article quality models and adopted by several research studies of Wikipedia article quality [e.g. 17, 22, 34, 41] was designed to address the first limitation by using article assessments at the time they were made as "ground truth." Article quality might drift in the periods between assessments, but it seems safe to assume that new quality assessments are accurate at the time they are made. A model trained on recent assessments can predict what quality label articles would receive if assessed in their current state.

[1] https://www.mediawiki.org/wiki/ORES

The main contribution of this paper is to introduce a method for constructing interpretable one-dimensional measures of article quality from Wikipedia quality assessments and the ORES article quality model. My approach improves upon ORES in two important ways. First, by using inverse probability weighting to calibrate the model, it is more accurate for typical research applications. Second, it does not depend on strong assumptions about the spacing between quality levels. In addition, this paper contributes an analysis of the performance of the ORES quality model that helps us understand the validity of previous work.

In §2, I provide a brief overview of quality measurement in peer production research, in which I foreground the importance of the assumptions needed to use machine learning predictions in downstream analysis—particularly the "evenly spaced" assumption used by Halfaker [16]—to justify the use of a handpicked weighted sum to combine article class probabilities. Next, in §3, I describe how to build accurate ordinal quality models that are appropriately calibrated for analyses of representative samples of Wikipedia articles or revisions. I also briefly explain how ordinal regression provides an interpretable one-dimensional measure of quality and how it relaxes the "evenly spaced" assumption. Finally, in §4 I present the results of my analysis to (1) show how the precision of the measurement depends on proper calibration and (2) demonstrate that the "evenly spaced" assumption is indeed violated. Despite this, I find that scores from the ordinal models are highly correlated with those from prior work, so the "evenly spaced" assumption may be acceptable in some applications. I conclude in §5 with recommendations for measuring article quality in future research.

2 BACKGROUND

Measurement is important to science as available knowledge often constrains the development of improved tools for advancing knowledge. For example, in the book *Inventing Temperature*, the philosopher and historian of science, Hasok Chang [11] documents how extending theories of heat beyond the range of human sense perception required scientists to develop new types of thermometers. This in turn required better knowledge of heat and of thermometric materials such as the freezing point of mercury. Part of the challenge of scientific advancement is that measurement devices developed under certain conditions may give unexpected results outside of the range in which they are calibrated: a thermometer will give impossibly low temperature readings when its mercury unexpectedly freezes. Today, machine learning models are used to extend the range of quality measurements in peer production research, but state of the art machine learning can be quite sensitive to the nuances of how their training data are selected [30].

2.1 Measuring Quality in Peer Production

As described in §1, measuring quality has been of great importance to peer production projects like Wikipedia and in the construction of knowledge about how such projects work. The foundation of article quality measurement in Wikipedia has been the peer production of article quality assessment organized by WikiProjects who develop criteria for articles in their domain [28]. This enables quality assessment to be consistent across different subject areas, but the procedures for assessing quality are tailored to the values of each WikiProject. Yet, like human sense perception of temperature, these quality assessments are limited in that they require human time and attention. Additionally, humans' limited ability to discriminate between levels on a scale limits the sensitivity of quality assessments. Articles are assessed irregularly and infrequently at the discretion of volunteer editors. Therefore, for most article revisions, it is not known what quality class the article would be assigned if it were newly assessed.

Researchers have proposed many ideas to extend the range of quality measurement beyond the direct perception of Wikipedians, including page length [7], persistent word revisions [1, 6], collaboration network structures [29], and template-based flaw detection [3]. Carefully constructed indexes benchmarked against English language Wikipedia quality assessments might allow quality measurement of articles that have not been assessed or in projects that have underproduced article assessments [24]. However, such indexes may lack emic validity if they fail to capture important aspects of quality or if notions of quality vary between linguistic communities and might even shape the editing activity in unexpected ways that could ultimately defeat their purpose [15, 35]. Peer produced quality labels depend on the limited capacity of volunteer communities to coordinate quality assessment, but also provide impressive validity for evaluating projects on their own terms.

2.2 Article Quality Models Extend Measurement to Unassessed Articles

Perhaps the most successful approaches to extending the range of quality measurements use machine learning models trained on available article quality assessments to predict the quality of revisions that have not been assessed. The ORES article quality model (henceforth ORES) implements this approach, but other similar article quality predictors have been developed [2, 12, 13, 29, 32, 44], and additional features including those based on language models can substantially improved classification performance compared to ORES [33]. The ORES model is a tree-based classifier that predicts the quality class of a Wikipedia article at the time it is assessed.[2] These tree-based models are reasonable for practical purposes with the reported ability to predict within 1 level of the true quality class with 90% accuracy (although in §4.2 I find a decline in accuracy in a more recent dataset). Yet, because these models do not account for the ordering of quality labels, the use of these predictions in downstream analysis introduces complicated methodological challenges.

The ORES classifiers are fit using sklearn through minimization of the multinomial deviance as shown [18, 27]:

$$L(y_i, p(x_i)) = -\sum_{k=1}^{K} I(y_i = \mathcal{G}_{i,k}) \log p_k(x_i) \quad (1)$$

For each article i with predictors x_i which has been labeled with a quality class y_i, the ORES model outputs an estimated probability $p_k(x_i)$ that the article belongs to each quality class $k \in \{Stub, Start, C\text{-}class, B\text{-}class, Good\ Article, Featured\ Article\}$. The predicted probabilities $p(x_i)$ sum to one so the ORES model outputs a unit vector for each article. If $\mathcal{G}_{i,k}$, the most probable quality class

[2]The system uses cross-validation to select among candidates that include random-forest and boosted decision tree models

(MPQC) according to the model, is the true label then $I(y_i = \mathcal{G}_{i,k})$ equals 1 (I is the indicator function) and the log predicted probability $p_k(x_i)$ of the correct class is subtracted from the loss $L(y_i, p(x_i))$. Note that this model does not use the fact that article quality classes are ordered. If it did, then it would have to penalize an incorrect classification of a *Good Article* as *C-class* more than a classification of a *Good Article* as *B-class*. In this model, different quality classes have no intrinsic rank or ordering and thus are akin to different categories of article subjects like animals, vegetables, or minerals.

The MPQC is perhaps the most natural way to use the ORES output to measure quality. It has been used in several studies including to provide evidence that politically polarized collaboration on Wikipedia leads to high quality articles [34] and to understand the relationship between article quality and donation [22]. However, the MPQC is limited in that it does not measure quality differences between articles that have the same MPQC. Consider two hypothetical articles; the first has the multinomial prediction $(0.1, 0.3, 0.4, 0.075, 0.075, 0)$ and the second has the prediction $(0.075, 0.075, 0.4, 0.3, 0.1, 0)$. The MPQC will assign both the *C-class* label even though the first article has an even chance at being a *Stub* or *Start-class* while the second article has an even chance at being a *B-class* or even *Good article*. At best, the MPQC has limited sensitivity to subtle variations or gradual changes in quality [16].

2.3 Combining Scores for Granular Measurement

To further extend the range of article quality measurement within article quality classes, Halfaker [16] constructed a numerical quality score using a linear combination (a weighted sum) of the elements of the multinomial prediction $p(x_i)$. This is advantageous from a statistical perspective as it naturally provides a continuous measure of quality which can typically justify a normal or log-normal statistical model. It can also support higher-order aggregations for measuring the quality of a set of articles [16]. Halfaker handpicks the coefficients $[0, 1, 2, 3, 4, 5]$ to make a linear combination of the predictions under the assumption "that the ordinal quality scale developed by Wikipedia editors is roughly cardinal and evenly spaced," which I refer to the "evenly spaced" assumption. It essentially says that a *Start-class* article has one more unit quality of a *Stub-class* article; that a *C-class* article has one more unit of quality than a *Start-class* article and so on. This approach is being adopted by other researchers including by Arazy et al. [4].

The considerable degree of effort and expertise required to raise articles to higher levels of quality raises doubt in the assumption [20]. Higher quality levels correspond to increasing completeness, encyclopedic character, usefulness to wider audiences, incorporation of multimedia, polished citations, and adherence to Wikipedia's policies. The English language Wikipedia editing guideline on content assessment[3] define a *Good Article* as "useful to nearly all readers, with no obvious problems" and a *Featured Article* article as "professional, outstanding and thorough." According to Wikipedians, it can take "three to six months of full time work" to write a featured article.[4] Are we to assume that the difference in quality between a *Good article* and a *Featured article* is measurably the same as that between a *Stub* defined as as "little more than a dictionary definition" and a *Start-class* article which is "a very basic description of the topic?" How could we even answer this question?

If the "evenly spaced" assumption is reasonable, then so is Halfaker's [16] weighted sum approach. But if increasing Wikipedia article classes do not represent roughly equal improvements in quality, then Halfaker's approach may not be accurate. Suppose that a *B-class* has not 1, but 2 units of quality greater than a *C-class* article, then Halfaker could have underestimated the improvement in the knowledge gap of women scientists, which was considerably driven by improvement in *B-class* articles. It turns out that a straightforward extension of the ORES article quality model based on ordinal regression can both relax the "evenly spaced" assumption and provide a better calibrated and more accurate one-dimensional measure of quality. In the next section, I describe my implementation of the approach.

3 DATA, METHODS AND MEASURES

My approach involves using Bayesian ordinal regression models that predict the labels using the ORES predicted probabilities for the quality classes to quantify the distance between quality levels. I provide a brief overview of ordinal regression as needed to explain my approach to measuring quality. Understanding ordinal regression depends on background knowledge of odds and generalized linear models. I recommend McElreath and Safari [25] for reference.

3.1 Bayesian Ordinal Regression

Ordinal regression predicts quality class membership using a single linear model and identifies boundaries between classes using the log cumulative odds link function shown below in Equation 2. The log cumulative odds is not the only possible choice of link function, but it is the most common, easiest to interpret, and is appropriate here.

$$\log \frac{\Pr(y_i \leq k)}{1 - \Pr(y_i \leq k)} = \alpha_k - \phi_i \qquad (2)$$
$$\phi_i = Bx_i$$

As in Equation 1 above, y_i is the quality label for article i. The left hand side of Equation 2 gives the log odds that y_i is less than or equal to quality level k. The ordinal quality measure is given by a linear model $\phi_i = Bx_i$ (x_i is a vector of transformed ORES scores for article i). Key to interpreting ϕ_i as a quality measure are the intercept parameters a_k for each quality level k. The log cumulative odds (the log odds that the article y_i has quality less than or equal to k) are given by the difference between the intercept and the linear model a_k - ϕ_i. Therefore, if $\phi_i = \alpha_k$ then the chances that $i <= k$ equal the chances that $i > k$. When ϕ_i is less than α_k, chances are that quality of article i is less than or equal to quality level k. As $\phi_i - \alpha_k$ increases so do the chances that article i is of quality better than k. In this way, the threshold parameters a_k define quantitative article quality levels on the scale of the ordinal quality measure ϕ_i.

[3] https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=1023695750

[4] Public statement by Stuart Yeates, an expert Wikipedian; quoted with permission. https://lists.wikimedia.org/hyperkitty/list/wiki-research-l@lists.wikimedia.org/message/7U35LHAXRWEPABN75DOTPOIEA2VYCTQQ/

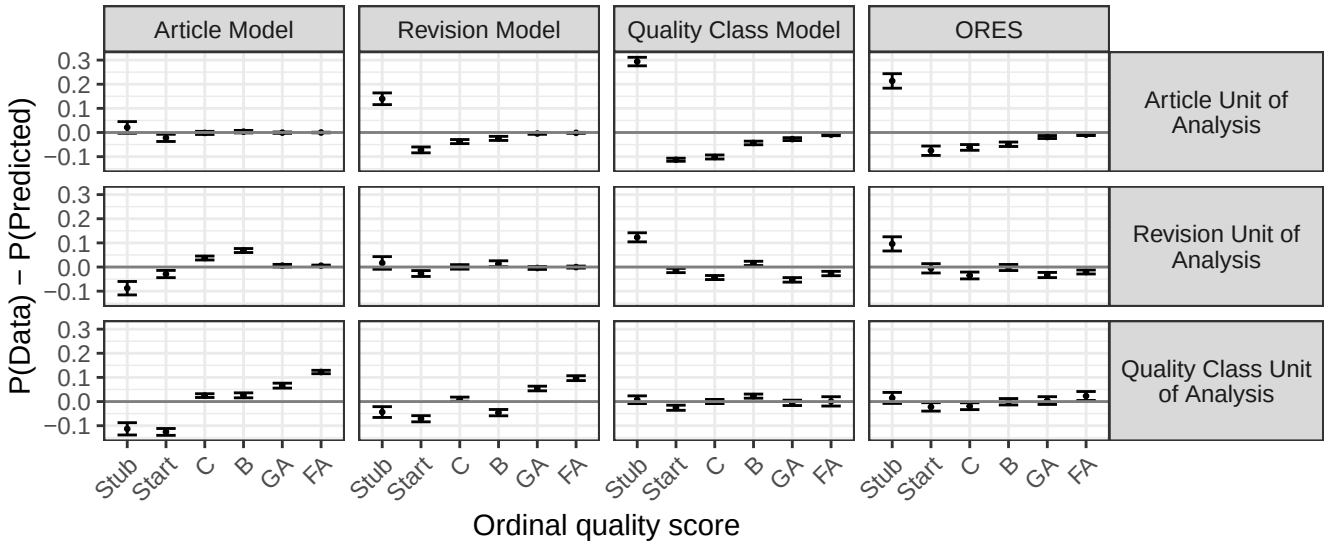

Figure 1: Calibration of each predictive quality model on datasets representative of each unit of analysis (article, revision, quality class). Each chart shows, for each quality class, the miscalibration of a model (columns) with respect to a dataset weighted to represent a unit of analysis (rows). The y-axis shows difference between the true probability of the quality class and the average predicted probability of that class, given a chosen unit of analysis. Points close to zero indicate good calibration. For example, the top-left chart shows that the article model is well-calibrated to the dataset on which it was fit and the middle-left chart shows that the article model predicts that articles are *Stubs* with probability greater than the frequency of *Stubs* in a random sample of revisions. Error bars show 95% confidence intervals.

Informally, an ordinal regression model maps a linear regression model to the ordinal scale using the log cumulative odds link function. It does this through the inference of thresholds that partition the range of linear predictions into regions defined by thresholds. When the linear predictor for an article crosses a threshold, the probability that the article has quality equal to or lesser than that corresponding to the threshold begins to decrease.

Bayesian inference allows interpreting model parameters like ϕ_i and α_k as random variables and provides accurate quantification of uncertainty in thresholds and predictions. I fit models using the R package Bayesian regression modeling using Stan (`brms`) [8] version 2.15.0. I use the default priors for ordinal regression, which are weakly informative. Due to the large sample size, these priors will be overwhelmed by the data and have little influence over results. I confirmed this by fitting equivalent frequentist models using the `polr` function in the `MASS` R package [40] and found that the estimates of intercepts and coefficients were very close.

The six quality scores output by the ORES article quality classifier are perfectly collinear by construction because they sum to one. This means they cannot all be included in the same regression model. Since interpreting the coefficients is not important, I take the linear transformation of the ORES scores using appropriately weighted principle component analysis (PCA) and use the first five principle components as the independent variables. This is simpler and more statistically efficient than a model selection procedure.

3.2 Dataset and Model Calibration

I draw a new random sample of 5,000 articles from each quality class to develop my models. I first reuse code the `articlequality`[5] Python package to process the March 2020 XML dumps for English Wikipedia and extract up-to-date article quality labels. I then select pages that have been assessed by a member of at least one WikiProject. Following prior work, if an article is assessed at different levels according to more than one WikiProject, I assign it to the highest such level and I drop articles having the rarely used quality class "A" [16, 41, 42]. Next, I use the `revscoring`[6] Python package to obtain the ORES quality scores at the time that they were labeled. Some of these revisions have been deleted leading to missing observations at each quality level. The number of articles sampled in each quality class are shown in Table 1. I reserve a random sample of 2000 articles which I use in reporting my results and fit my ordinal regression models on the remainder.

The ORES article quality classifiers are fit on a "balanced" dataset having an equal number of articles in each quality class. Constructing a balanced dataset by oversampling or undersampling is a common practice in machine learning for probabilistic loss functions such as the multinomial log-loss because models fit on highly imbalanced datasets will simply learn to predict the most common class. An ORES quality score is the probability that an article is a member of a quality class under the assumption that the article

[5] https://pypi.org/project/articlequality/
[6] https://pypi.org/project/revscoring/

Table 1: Number of articles sampled at each quality level.

Label	No. Articles	No. Revisions	Sample Size	Article Weights	Revision Weights
Stub	3,359,351	12,005,611	4,969	4.23	2.52
Start	1,019,038	7,828,335	4,979	1.28	1.64
C	235,655	3,889,639	4,988	0.30	0.81
B	128,875	3,640,591	4,990	0.16	0.76
GA	31,808	924,468	4,999	0.04	0.19
FA	7,438	365,255	4,995	0.01	0.08

was drawn from a population where each quality class contains an equal number of articles. Simply put, the model has learned that each quality class is about the same size.

This, of course, is not representative of the overall article quality on Wikipedia, which is highly skewed with over 3 million *Stubs* but only around *7,000 Featured* articles as shown in Table 1. Although using a balanced dataset likely improves the accuracy of the ORES models, for the ordinal regression models, the choice of unit of analysis presents a trade-off between accuracy in a representative sample of articles or revisions and accuracy within each quality class.

The "balanced" dataset corresponds to the *quality class* unit of analysis because each quality class has equal representation. However, researchers are more interested in analyzing representative samples of *articles* of article *revisions*. For example, the article unit of analysis would be used to estimate the average quality of a random sample of articles and the revision unit of analysis might be used to model the change in the quality of an encyclopedia over time.

Although oversampling or undersampling is a common practice in machine learning because it can improve predictive performance, it can lead to badly calibrated predictive probabilities as shown in Figure 1. Calibration means that, on average, the predicted probability of a quality class equals the average true probability of that class for the unit of analysis. Weighting allows the use of the balanced dataset to estimate a model as if the dataset were a uniform random sample of a different population. My method uses inverse probability weighting to fit ordinal regression models at each of these three levels of analysis so that the probability estimates are well calibrated. For example, the model calibrated to the article unit of analysis has weights corresponding to each quality class calculated as the probability of the quality class over all articles divided by the probability of the quality class in the sample. The size of the sample and the weights for the article and revision levels of analysis are also shown in Table 1.

4 RESULTS

I first report my findings about the spacing of the quality classes in each of the models in §4.1. Quality classes are not evenly spaced, especially when articles or revisions are the unit of analysis. Next, in §4.2 I report the accuracy of each of the models and the uncertainty of the ordinal quality scale. All models perform similarly to or better than the MPQC within the pertinent unit of analysis. The unweighted model provides the best accuracy and lowest uncertainty across the entire range of quality levels, but is poorly calibrated for other units of analysis. Finally, in §4.3 I show that all quality measures are highly correlated, but the ordinal quality measures agree with one another more than with the "evenly spaced" measure.

4.1 Spacing of Quality Classes

The grid of charts in Figure 2 shows quality scores and thresholds for each model (columns) and article quality level (rows). Each chart shows the histogram of quality scores ϕ_i given to articles having the true quality label corresponding to the row of the grid. The histograms are colored to indicate regions where the model correctly predicts that articles belong to their true class. Vertical dashed lines show the thresholds inferred by the model with 95% credible intervals colored in gray. Different models have different scales of scores, so Figure 2 shows results normalized between 0 and 1.

No matter the unit of analysis, article quality classes are not evenly spaced. The quality class model provides a quality scale in which *Featured* articles take up 27% of the scale and are expected to score in the range of $[0.73, 1]$, but probable *C-class* articles only span 14% of the scale in the range $[0.31, 0.45]$. Researchers are likely to be interested models calibrated to the article or revision units of analysis and in these cases the quality classes are far from evenly spaced. The *revision model* assigns 28% of the scale to *Stubs*, from 0 to 0.28. It assigns *C-class* articles the smallest part of the scale only 4% of it, from 0.54 to 0.58. The *article model* is even more extreme. It assigns *Stubs* to the interval $[0, 0.39]$, 39% of the scale and the space between thresholds defining the range of *C-class* articles is so narrow that it virtually never predicts that an article will be C-class. In general terms, the *quality class model* gives more equal amounts of space to each quality class compared to the other models, while reserving nearly the top half of the scale for the top 2 quality classes. The *revision model* and *article model* do the opposite and use the bottom half of the scale to account for differences within the bottom two quality classes, leave some room for *B-class* articles, but squeeze the top end of the scale and *C-class* articles in relatively small intervals.

4.2 Accuracy and Uncertainty

I evaluate the predictive performance in terms of *accuracy*, the proportion of predictions of article quality that are correct. To allow comparison with the reported accuracy of the ORES quality models, I also report *off-by-one accuracy*, which includes predictions within one level of the true quality class among correct predictions.

As shown in Table 2, the ordinal regression models have better predictive ability than the MPQC except when the unit of analysis is the quality class. In this case, the best ordinal quality model

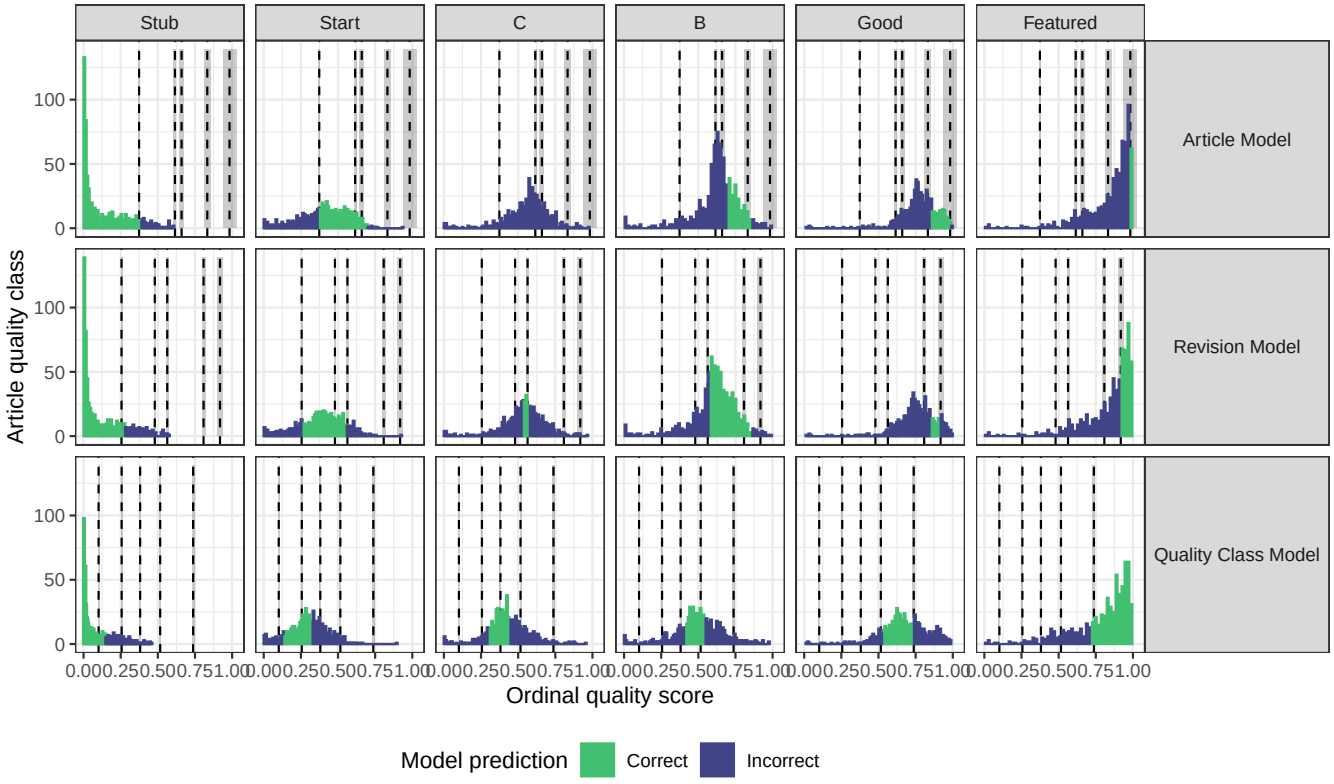

Figure 2: Quality scores and predictions of the ordinal regression models. Rows in the grid of charts correspond to the ordinal quality models calibrated to the indicated unit of analysis and columns correspond to sampled articles having the indicated level of quality as assessed by Wikipedians. Each chart shows the histogram of scores, thresholds inferred by the ordinal model with 95% credible intervals colored in gray, and colors indicating when the model makes correct or incorrect predictions. The weighted model has better accuracy for the lowest quality classes where a vast majority of Wikipedia articles are located, but poor accuracy in the higher quality classes where the unweighted model is good. The thresholds are not evenly spaced, especially in *revision model* and *article model* which put more weight on lower quality classes and infer that the gaps between *Stub* and *Start* and between *Start* and *C-class* articles are considerably wider than the gap between *C-class* and *B-class* articles.

has worse accuracy than the MPQC but slightly better off-by-one accuracy. Table 2 shows accuracy and off-by-one accuracy weighted for each unit of analysis. Accuracy for a given unit of analysis depends on having a model fit to data representative of that unit of analysis. Accuracy scores are higher when greater weight is placed on lower article quality classes, suggesting that it is easier to discriminate between these classes.

The ORES article quality model has been quickly adopted by researchers, but its accuracy is limited. While off-by-one accuracy is above 90% when the article is the unit of analysis, the MPQC only predicts the correct quality class 55% of the time when the quality class is the unit of analysis.

The trade-offs in selecting a unit of analysis on which to calibrate the models are further illustrated by Figure 3, which plots the size of the 95% credible intervals as a function of the quality scores for each model. As in Figure 2, quality scores in this plot are rescaled between 0 and 1. The models calibrated to articles or revisions have more certainty in the lower range of the quality scale compared to the model that places equal weight in all quality classes. This comes with a trade-off for the higher range of quality. Whereas the *quality class model* has relatively low uncertainty across the entire range of quality, the *revision model* and *article model* have greater uncertainty at higher levels of quality.

4.3 Correlation Between Scores

Although the models have different predictive performances and uncertainties, as measures of quality, they are nearly perfectly correlated with one another as shown in Figure 4. For each quality score, including Halfaker's [16] "evenly spaced" weighted sum, Figure 4 shows a scatter plot and two correlation statistics: Kendall's τ and Pearson's r. Pearson's r is the standard linear correlation coefficient and Kendall's τ is a nonparametric rank-based correlation defined as the probability that the quality scores will agree about which of

Table 2: Accuracy of quality prediction models depends on the unit of analysis. The greatest accuracy and off-by-one accuracy scores are highlighted. Models are more accurate when calibrated on the same unit of analysis on which they are evaluated. Compared to the MPQC, the ordinal quality models have better accuracy when revisions or articles are the unit of analysis. When the quality class is the unit of analysis, the ordinal quality model has worse accuracy, but predicts within 1 quality class with slightly better accuracy.

Unit of analysis	Model	Ordinal Model?	Accuracy	Off-by-one Accuracy
No weights	Article Model	Yes	0.33	0.75
No weights	Revision Model	Yes	0.44	0.84
No weights	Quality Class Model	Yes	0.52	0.87
No weights	ORES MPQC	No	0.55	0.86
Revision Weights	Article Model	Yes	0.57	0.87
Revision Weights	Revision Model	Yes	0.61	0.92
Revision Weights	Quality Class Model	Yes	0.54	0.88
Revision Weights	ORES MPQC	No	0.58	0.9
Article Weights	Article Model	Yes	0.76	0.97
Article Weights	Revision Model	Yes	0.73	0.96
Article Weights	Quality Class Model	Yes	0.63	0.92
Article Weights	ORES MPQC	No	0.65	0.94

any two articles has higher quality minus the probability that they will disagree.

According to Pearson's r all the quality scores are highly correlated with correlation coefficients of about 0.98 or higher. Kendall's τ measures nonlinear correlation and reveals discrepancies between the ordinal models and the "evenly spaced" measures. The Pearson correlation between scores from the *revision model* and the scores from the *quality class model* are about the same as the correlation between the *revision model* scores and the "evenly spaced" scores ($r = 0.98$). However, according to Kendall's τ, scores from the *revision model* are more similar to those from the *quality class model* ($r = 0.98$) than to the scores from the "evenly spaced" approach ($r = 0.9$).p

The evenly spaced model is more likely to disagree with the model-based scores than any of the model-based scores are to disagree with one another as visualized in the scatter plots in Figure 4. Disagreement between the "evenly spaced" method and the ordinal models is greatest among articles in the middle of the quality range.

5 DISCUSSION

Past efforts to extend the measurement of Wikipedia article quality from peer-produced article quality assessments to unassessed versions of articles and from the discrete to the continuous domain have relied upon machine learning and expedient but untested assumptions like that quality levels are "evenly spaced."

While I suggest technical improvements for statistical models for measuring quality, I also find that scores from my models are highly correlated to those obtained under the "evenly spaced" assumption. I set out to provide a better way to convert the probability vector output by the ORES article quality model into a continuous scale and to test the assumption that the quality levels are evenly spaced. I used ordinal regression models to infer spacing between quality levels and used the linear predictor of these models as a continuous measure of quality. While I found in §4.1 that the quality levels are not evenly spaced and that the spacing depends on the unit of analysis to which the models are calibrated, I also showed in §4.3 that the model-based quality measures are highly, although not perfectly, correlated with Halfaker's [16] "evenly spaced" measure. This provides some assurance that past results built on this measure are unlikely to mislead. That said, I recommend that future work adopt appropriately calibrated model-based quality measures instead of the "evenly spaced" approach, and I argue that it is important to improve the accuracy of article quality predictors to enable more precise article quality measurement.

5.1 Recommendations for Measuring Article Quality

How should future researchers approach the question of how to measure Wikipedia article quality? While I cannot provide a final or complete answer to the question, I believe the exercise reported in this paper provides some insights on which to base recommendations. It is important to note that I consider here only approaches to measuring quality that assume the use of a good predictor of article quality assessment, such as the ORES quality model. I do not consider other based approaches such as those based on indexes [24] described in §2.

5.1.1 Use the principle components of ORES scores for statistical control of article quality. In many statistical analysis, the only purpose of measuring quality will be as a statistical control or adjustment. For example, Zhang et al. [43] used the MPQC as a control variable in a propensity score matching analysis of promotion to *Featured* article status, but as argued in §3, the MPQC provides less information than the vector of ORES scores. Using the principle components is simpler than using an ordinal quality model. I recommend obtaining ORES scores for your dataset, taking the principle components, and dropping the least significant one to remove collinearity.

5.1.2 Use ordinal quality scores when article quality is an independent variable. In other cases, research questions will ask how article quality is related to an outcome of interest, like how Kocielnik et al. [22] set out to explore factors associated with donations to the Wikimedia Foundation. They use the MPQC as an independent

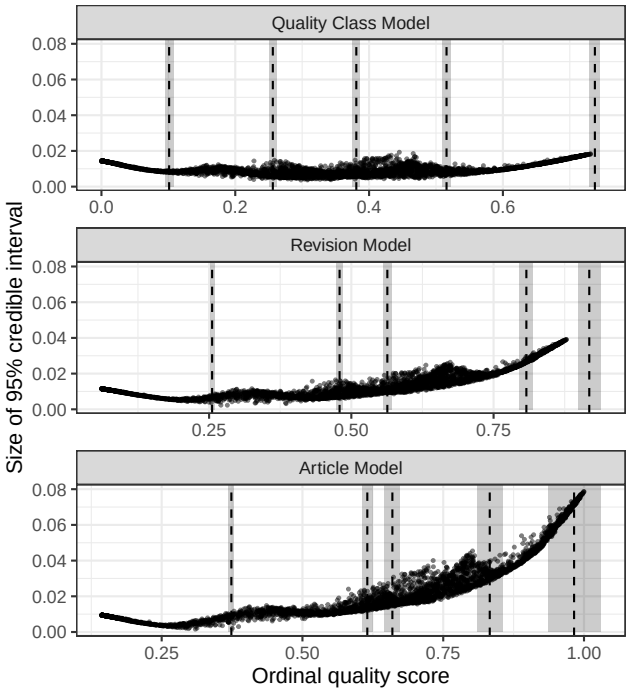

Figure 3: Uncertainty in ordinal quality scores for models calibrated at each unit of analysis. Points show the size of the 95% credible interval for the ordinal quality score for each article in the dataset. Models fit on a "balanced" dataset have low uncertainty across the range of quality; models calibrated to the revision and article levels of analysis have less uncertainty at the low end of the quality scale, but greater uncertainty at the higher end of the scale.

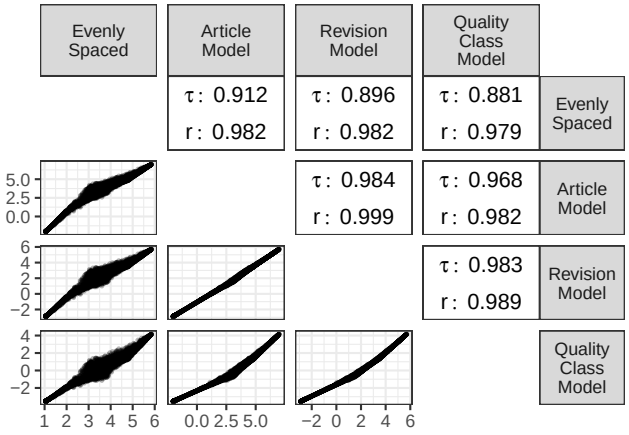

Figure 4: Correlations between quality measures show that the different approaches to measuring quality are quite similar. "Weighted sum" refers to Halfaker's [16] measure using a weighted sum with handpicked coefficients under the "evenly spaced" assumption. Lower values of Kendall's τ, a nonparametric rank correlation statistic compared to Pearson's r suggest nonlinear differences between the weighted sum and the other measures.

variable, which complicates their analysis. Although they conclude that "pages with higher quality attract more donations," this is not strictly true. They actually found a nonlinear relationship where readers of *B-class* articles were more likely to donate than readers of *Featured* articles. Using a continuous measure of quality is more convenient when the average linear relationship is the target of inference.

I recommend using an ordinal regression model appropriate to the downstream unit of analysis because this will justify the interpretation of the measure. If the downstream unit of analysis differs substantively from those used here, such as if different selection criteria are applied, I recommend reusing my code to calibrate a new ordinal regression model to a new dataset. Otherwise, reusing one of my models should be adequate. Code, data and instructions for replicating or reusing this analysis are available at [link removed for anonymous review]. Finally, in the Bayesian framework, the scores are interpretable as random variables. This provides a justification for incorporating the variance of these scores as measurement errors to improve estimation in downstream analysis [25].

5.1.3 Use the MPQC or ordinal quality scores when article quality is the dependent variable. Using the MPQC as the outcome in an ordinal regression model, as is done by Shi et al. [34] in their analysis of Wikipedia articles with politically polarized editors is a reasonable choice as long as it provides sufficient variation and a more granular quality measure is not needed. Although it is theoretically possible that using the MPQC might introduce statistical bias because it less accurate than ordinal quality scores for units of analysis other than the quality class and omits variation within quality classes, such threats to validity do not seem more significant than the threat introduced by inaccurate predictions. If the MPQC does not provide sufficient granularity and a continuous measure is desired as in Halfaker [16] or Arazy et al. [4], I recommend using a measure based on ordinal regression as described in §5.1.2 above.

5.2 Limitations

Although intuitions about the varying degrees of effort required to develop articles with different levels of quality led me to question the "evenly spaced" assumption, my findings that quality classes are not evenly spaced do not reflect relative degrees of effort. Rather, spaces between levels are chosen to link a linear model to ordinal data. The spacing of intervals depends on the ability of the ORES scores to predict quality classes. The ORES article quality model has relative difficulty classifying *C-class* and *B-class* articles [16]. Perhaps the differences between these quality classes are minor compared to the other classes. Maybe ORES lacks the features or ability to model these differences and the space between these classes will grow if its predictive performance improves.

The usefulness of article quality scores depends on the accuracy of the model. The ORES quality models are accurate enough to be

useful for researchers, but they still they only predict the correct quality class 55% of the time on a balanced dataset. Of course, this limits the accuracy of the ordinal regression models reported here. Furthermore, while the ORES quality models were designed with carefully chosen features intended to limit biases [17], it is still quite plausible that the accuracy of predictive quality models may vary depending on characteristics of the article [21]. Such inaccuracies may introduce bias, threaten downstream analysis or lead to unanticipated consequences of collaboration tools built upon the models [37]. Therefore, improving the accuracy of article quality prediction models is important to the reliability of future article quality research. Adopting machine learning models that can incorporate ordinal loss functions is a promising direction and can reduce the need for auxiliary ordinal regression models [9].

This paper only considers measuring article quality for English language Wikipedia, but expanding knowledge of collaborative encyclopedia production depends on studying other languages as audiences and collaborative dynamics can vary greatly between projects [19, 23, 36]. Other languages carry out quality assessments [24], and some of these have been used to build ORES article quality models. Future work should extend this project to provide multilingual article quality measures in one continuous dimension.

An additional limitation stems from the likelihood that peer-produced quality labels are biased. For instance, the English Wikipedia community has a well-documented pattern of discrimination against content associated with marginalized groups such as biographies of women [26, 38] and indigenous knowledge [39]. Although demonstrating biases in article quality assessment is a task for future research, if Wikipedians' assessments of article quality are biased then model predictions of quality will almost certainly be as well.

6 CONCLUSION

Measuring article quality in one continuous dimension is a valuable tool for studying the peer production of information goods because it provides granularity and is amenable to statistical analysis. Prior approaches extended ORES article quality prediction into a continuous measure under the "evenly spaced" assumption. I showed how to use ordinal regression models to transform the ORES predictions into a continuous measure of quality that is interpretable as a probability distribution over article quality levels; provides an account of its own uncertainty and does not assume that quality levels are "evenly spaced." Calibrating the models to the chosen unit of analysis improves accuracy for research applications. I recommend that future work adopt this approach when article quality is an independent variable in a statistical analysis.

REFERENCES

[1] B. Thomas Adler and Luca de Alfaro. 2007. A Content-Driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 261–270.

[2] Maik Anderka and Benno Stein. 2012. A Breakdown of Quality Flaws in Wikipedia. In *Proceedings of the 2Nd Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality '12)*. ACM, New York, NY, 11–18.

[3] Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting Quality Flaws in User-Generated Content: The Case of Wikipedia. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 981–990.

[4] Ofer Arazy, Aron Lindberg, Mostafa Rezaei, and Michele Samorani. 2019. The Evolutionary Trajectories of Peer-Produced Artifacts: Group Composition, the Trajectories' Exploration, and the Quality of Artifacts. *MIS Quarterly* (Dec. 2019).

[5] Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works and How You Can Be a Part of It.* No Starch Press, San Francisco, CA.

[6] Susan Biancani. 2014. Measuring the Quality of Edits to Wikipedia. In *Proceedings of The International Symposium on Open Collaboration (OpenSym '14)*. ACM, New York, NY, USA, 33:1–33:3.

[7] Joshua E. Blumenstock. 2008. Size Matters: Word Count as a Measure of Quality on Wikipedia. In *Proceeding of the 17th International Conference on World Wide Web - WWW '08*. ACM Press, Beijing, China, 1095.

[8] Paul-Christian Bürkner. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80, 1 (Aug. 2017), 1–28.

[9] Jaime S Cardoso, Jaime Cardoso, and Inescporto Pt. 2007. Learning to Classify Ordinal Data: The Data Replication Method. *Journal of Machine Learning Research* 8 (2007), 37.

[10] Kaylea Champion and Benjamin Mako Hill. 2021. Underproduction: An Approach for Measuring Risk in Open Source Software. *IEEE International Conference on Software Analysis, Evolution and Reengineering* (Feb. 2021). arXiv:2103.00352 [cs.SE]

[11] Hasok Chang. 2004. *Inventing Temperature.* OUP, Oxford.

[12] Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality Assessment of Wikipedia Articles Without Feature Engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, New York, NY, USA, 27–30.

[13] Gregory Druck, Gerome Miklau, and Andrew McCallum. 2008. Learning to Predict the Quality of Contributions to Wikipedia. In *WikiAI*. 6.

[14] Andrea Forte and Amy Bruckman. 2005. Why Do People Write for Wikipedia? Incentives to Contribute to Open-Content Publishing. In *Proceedings of GROUP*. 6.

[15] C. A. E. Goodhart. 1984. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice: The UK Experience*, C. A. E. Goodhart (Ed.). Macmillan Education UK, London, 91–121.

[16] Aaron Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *Proceedings of the 13th International Symposium on Open Collaboration (OpenSym '17)*. Association for Computing Machinery, New York, NY, USA, 1–9.

[17] Aaron Halfaker and R Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. 4, 148 (Oct. 2020), 37.

[18] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. 2018. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

[19] Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, Atlanta, Georgia, USA, 291–300.

[20] Dariusz Jemielniak. 2014. *Common Knowledge?: An Ethnography of Wikipedia.* Stanford University Press.

[21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]* (Sept. 2016). arXiv:1609.05807 [cs, stat]

[22] Rafal Kocielnik, Os Keyes, Jonathan T. Morgan, Dario Taraborelli, David W. McDonald, and Gary Hsieh. 2018. Reciprocity and Donation: How Article Topic, Quality and Dwell Time Predict Banner Donation on Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–20.

[23] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. ACM, New York, NY, USA, 618–626.

[24] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics* 4, 4 (Dec. 2017), 43.

[25] Richard McElreath and an O'Reilly Media Company Safari. 2018. *Statistical Rethinking.*

[26] Amanda Menking, Ingrid Erickson, and Wanda Pratt. 2019. People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830.

[28] Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works.* No Starch Press.

[29] Narun Raman, Nathaniel Sauerberg, Jonah Fisher, and Sneha Narayan. 2020. Classifying Wikipedia Article Quality With Revision History Networks. In *Proceedings of the 16th International Symposium on Open Collaboration (OpenSym 2020)*. Association for Computing Machinery, New York, NY, USA, 1–7.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *International Conference on Machine Learning*. PMLR, 5389–5400.

[31] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2021. A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). *arXiv:2008.12314 [cs]* (Jan. 2021). arXiv:2008.12314 [cs]

[32] Soumya Sarkar, Bhanu Prakash Reddy, Sandipan Sikdar, and Animesh Mukherjee. 2019. StRE: Self Attentive Edit Quality Prediction in Wikipedia. *arXiv:1906.04678 [cs]* (June 2019). arXiv:1906.04678 [cs]

[33] Manuel Schmidt and Eva Zangerle. 2019. Article Quality Classification on Wikipedia: Introducing Document Embeddings and Content Features. In *Proceedings of the 15th International Symposium on Open Collaboration (OpenSym '19)*. Association for Computing Machinery, New York, NY, USA, 1–8.

[34] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A. Evans. 2019. The Wisdom of Polarized Crowds. *Nature Human Behaviour* 3, 4 (April 2019), 329–336.

[35] Marilyn Strathern. 1997. 'Improving Ratings': Audit in the British University System. *European Review* 5, 3 (July 1997), 305–321.

[36] Nathan TeBlunthuis, Tilman Bayer, and Olga Vasileva. 2019. Dwelling on Wikipedia: Investigating Time Spent by Global Encyclopedia Readers. In *OpenSym '19, The 15th International Symposium on Open Collaboration*. Skövde, Sweden, 14.

[37] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2021. Effects of Algorithmic Flagging on Fairness: Quasi-Experimental Evidence from Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 56:1–56:27. arXiv:2006.03121

[38] Francesca Tripodi. 2021. Ms. Categorized: Gender, Notability, and Inequality on Wikipedia. *New Media & Society* (June 2021), 14614448211023772.

[39] Maja van der Velden. 2013. Decentering Design: Wikipedia and Indigenous Knowledge. *International Journal of Human–Computer Interaction* 29, 4 (March 2013), 308–316.

[40] W. N Venables, Brian D Ripley, and W. N Venables. 2002. *Modern Applied Statistics with S*. Springer, New York.

[41] Morten Warncke-Wang, Vladislav R. Ayukaev, Brent Hecht, and Loren G. Terveen. 2015. The Success and Failure of Quality Improvement Projects in Peer Production Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 743–756.

[42] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell Me More: An Actionable Quality Model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13)*. Association for Computing Machinery, New York, NY, USA, 1–10.

[43] Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel P. Robert, and Daniel M. Romero. 2017. Crowd Development: The Interplay between Crowd Evaluation and Collaborative Dynamics in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–21.

[44] Shiyue Zhang, Zheng Hu, Chunhong Zhang, and Ke Yu. 2018. History-Based Article Quality Assessment on Wikipedia. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 1–8.