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. Wikipedia community members and academic researchers have used article quality ratings for purposes like tracking knowledge gaps and studying how political polarization shapes collaboration. Even so, measuring quality presents many methodological challenges. The most widely used systems use quality assessments on discrete ordinal scales, 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 degrees of effort needed to raise Wikipedia articles to different quality levels. 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.

1 INTRODUCTION
Measuring content quality in peer production projects like Wikipedia is important so projects can learn about themselves and track progress. Measuring quality also helps build confidence that information is accurate and supports monitoring how well an encyclopedia includes diverse subject areas to identify gaps needing attention [31]. Measuring quality enables tracking and evaluating the progress of subprojects and initiatives organized to fill the gaps [16, 41]. Raising an article to a high standard of quality is a recognized achievement among contributors, so assessing quality can help motivate contributions [3, 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 [10].

Measuring quality also presents methodological and ontological challenges. How can “quality” be conceptualized so that measurement of the goals of a project and the value it produces can be precise and accurate? Language editions of Wikipedia, including English, peer produce quality labels 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 violates typical assumptions in statistical analysis. 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 Service1 (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 an article would receive if assessed in its current state.

This paper introduces a method for constructing interpretable one-dimensional measures of article quality from Wikipedia quality assessments and the ORES article quality model. The method improves upon prior approaches in two important ways. First, by using inverse probability weighting to calibrate the model, it is more accurate for typical research applications, and second, it does not depend on the assumption that quality levels are “evenly spaced;” which threatens the validity of prior research [4, 16]. In addition,

this paper helps us understand the validity of previous work by analyzing the performance of the ORES quality model and testing the "evenly spaced" assumption.

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 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, Hasok Chang [11], the philosopher and historian of science, 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. In addition, 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, such as page length [7], persistent word revisions [1, 6], collaboration network structures [29], and template-based flaw detection [5]. 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 improve 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. These tree-based models are reasonable for practical purposes with the reported ability to predict within one 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, since 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 scikit-learn through minimization of the multinomial deviance as shown [18, 27]:

\[
L(y_i, p(x_i)) = -K \sum_{k=1}^{K} I(y_i = k) \log p_k(x_i)
\]

For each article \(i\) with predictors \(x_i\) that 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 \(k \in \{\text{stub}, \text{start}, \text{C-class}, \text{B-class}, \text{Good article (GA)}, \text{Featured article (FA)}\}\). The predicted probabilities \(p_k(x_i)\) sum to one so the ORES model outputs a unit vector for each article. If \(\hat{G}_{i,k}\) is the most probable quality class (MPQC) according to the model, is the true label, then \(L(y_i = \hat{G}_{i,k}) = 1\ (I\ is\ the\ indicator\ function)\ and\ the\ log\ predicted\ probability\ \log p_k(x_i)\ of\ the\ correct\ class\ is\ subtracted\ from\ the\ loss\ \log L(y_i, p(x_i))\). Note that this model does not use the fact that

---

1The system uses cross-validation to select among candidates that include random-forest and boosted decision tree models.

2https://scikit-learn.org/stable/
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 the same chance at being a Stub or Start-class as the second article’s chance at being a B-class or even a 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 $p(x)$. 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, and that a C-class article has one more unit quality than a Start-class article and so on. This approach is being adopted by other researchers including Arazoy et al. [4].

The considerable degree of effort and expertise required to raise articles to higher levels of quality raises doubt in the assumption that the “evenly spaced” assumption is reasonable, then Halfaker’s weighted sum approach is too. But if increasing Wikipedia article classes do not represent roughly equal improvements in quality, this may threaten the accuracy of analysis dependent on the assumption. Suppose that a B-class article 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. In the next section, I provide a straightforward extension of the ORES article quality model based on ordinal regression that both relaxes the “evenly spaced” assumption and provides a better calibrated and more accurate one-dimensional measure of quality.

3 DATA, METHODS AND MEASURES

I use Bayesian ordinal regression models that use the ORES predicted probabilities to predict the quality class labels and quantify the distance between quality classes. I now 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 for all classes and identifies boundaries between classes using the log cumulative odds link function shown below in Eq. 2. The log cumulative odds is not the only possible choice of link function, but it is the most common, is the easiest to interpret, and is appropriate here.

$$\log \frac{Pr(y_i \leq k)}{1 - Pr(y_i \leq k)} = a_k - \phi_i \quad (2)$$

As in Eq. 1, $y_i$ is the quality label for article $i$. The left hand side of Eq. 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 $\phi_i = b x_i$ ($x_i$ is a vector of transformed ORES scores for article $i$). Key to interpreting $\phi_i$ 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 $a_k - \phi_i$. Therefore, if $\phi_i = a_k$ then the chances that $i < k$ equal the chances that $i > k$. When $\phi_i$ is less than or equal to quality level $k$. As $\phi_i - a_k$ 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 $\phi_i$.

Informally, an ordinal regression model maps a linear regression model to the ordinal scale using the log cumulative odds link function. It does this by inferring thresholds that partition the range of linear predictions. When the linear predictor for an article crosses a threshold, the probability that the article has quality greater than that corresponding to the threshold begins to increase.
Table 1: Numbers of articles and revisions, sample sizes, and regression weights for each quality level.

<table>
<thead>
<tr>
<th>Label</th>
<th>No. of articles</th>
<th>No. of revisions</th>
<th>Sample size</th>
<th>Article weights</th>
<th>Revision weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stub</td>
<td>3,359,351</td>
<td>12,005,611</td>
<td>4,969</td>
<td>4.23</td>
<td>2.52</td>
</tr>
<tr>
<td>Start</td>
<td>1,019,038</td>
<td>7,828,335</td>
<td>4,979</td>
<td>1.28</td>
<td>1.64</td>
</tr>
<tr>
<td>C</td>
<td>235,655</td>
<td>3,889,639</td>
<td>4,988</td>
<td>0.30</td>
<td>0.81</td>
</tr>
<tr>
<td>B</td>
<td>128,875</td>
<td>3,640,591</td>
<td>4,990</td>
<td>0.16</td>
<td>0.76</td>
</tr>
<tr>
<td>GA</td>
<td>31,808</td>
<td>924,468</td>
<td>4,999</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>FA</td>
<td>7,438</td>
<td>365,255</td>
<td>4,995</td>
<td>0.01</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Bayesian inference allows interpreting model parameters like $\phi_1$ and $\alpha_k$ 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, the data overwhelm the priors and the priors 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 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 from the article-quality\(^6\) 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 A-class quality level [16, 41, 42]. Next, I use the revscoring\(^7\) Python package to obtain the ORES scores of the labeled article versions. Some of these versions have been deleted leading to missing observations at each quality level. Table 1 shows the number of articles sampled in each quality class. I reserve a random sample of 2000 articles which I use in reporting my results and fit my ordinal regression models on the remainder.

---

\(^6\)https://pypi.org/project/articlequality (https://perma.cc/HR4H-MAZ9)

\(^7\)https://pypi.org/project/revscoring (https://perma.cc/HFN-V232)
The ORES article quality classifiers are fit on a “balanced” dataset having an equal number of articles in each quality class. Thus, an ORES score is the probability that an article is a member of a quality class under the assumption that the article was drawn from a population where each quality class contains an equal number of articles. Simply put, the model has learned from its training data that each quality class is about the same size.

This 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. Constructing a balanced dataset by oversampling is a common practice in machine learning because it can improve predictive performance. However, oversampling can also lead to badly calibrated predictive probabilities as shown in Fig. 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.

The “balanced” dataset on which ORES is trained has the quality class unit of analysis because each quality class has equal representation. However, researchers are more interested in analyzing representative samples of articles or 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. Weighting allows the use of the balanced dataset to estimate a model as if the dataset were a uniform random sample of a different unit of analysis. My method uses a balanced dataset to fit ordinal regression models with inverse probability weighting to calibrate each model to the unit of analysis of a research project. For example, each article in the model calibrated to the article unit of analysis is weighted by the probability of its quality class in the population of articles divided by the probability of its 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
which

The grid of charts in Fig. 2 shows quality scores and thresholds

1.

which are far from evenly spaced. The quality class model provides a quality scale in

ranges of scores, so Fig. 2 shows results normalized between 0 and 1. The models calibrated to article or revisions have less uncer-

tainty across the range of quality. Models calibrated to the other units of analysis have less uncertainty at the lower end of the scale, but greater uncertainty at the higher end of the scale.

4.1 Spacing of Quality Classes

The grid of charts in Fig. 2 shows quality scores and thresholds for each model (columns) and article quality level (rows). Each chart shows the histogram of quality scores \( \phi_i \) 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 ranges of scores, so Fig. 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 in 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 relatively 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 into relatively small intervals.

4.2 Accuracy and Uncertainty

I evaluate 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 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 Fig. 3, which plots the size of the 95% credible intervals as a function of the quality scores for each model. As in Fig. 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. While 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 Fig. 4. For each quality score, including the “evenly spaced” weighted sum, Fig. 4 shows a scatter plot and two correlation statistics: Kendall’s \( r \) and Pearson’s \( r \). Pearson’s \( r \) is the standard linear correlation coefficient and Kendall’s \( r \) is a nonparametric rank-based correlation defined as the probability that the quality scores will agree about which of

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. The quality class model has 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.
When the quality class is the unit of analysis, the ordinal quality model has worse accuracy, but predicts within one quality class with slightly better accuracy. 

According to Pearson’s $r$, all the quality scores are highly correlated with correlation coefficients of about 0.98 or higher. Kendall’s $\tau$ 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 $\tau$, 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 ($r = 0.9$). 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 Fig. 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 the “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 analyses, 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
“Evenly spaced” uses a weighted sum of the ORES scores with handpicked coefficients [16]. Lower values of Kendall’s $W$ indicate that Wikipedia articles with politically polarized editors, is a reasonable approximation of the quality of the article. Although they use the MPQC as an independent variable, which complicates their analysis. Although they conclude that the quality class and omits variation within quality classes, such a measure is not needed. Although it is theoretically possible that using the MPQC might introduce statistical bias because it less accurately captures the measure's ability to model these differences and the space between these classes will grow if its predictive performance improves.

### 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 necessarily 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 ordinal regression model 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 or is not able to improve model prediction because the differences 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 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 of results.

Figure 4: Correlations between quality measures show that the different approaches to measuring quality are quite similar. "Evenly spaced" uses a weighted sum of the ORES scores with handpicked coefficients [16]. Lower values of Kendall’s $W$, a nonparametric rank correlation statistic, compared to Pearson’s $r$ suggest nonlinear differences between the weighted sum and the other measures.

<table>
<thead>
<tr>
<th>Evenly spaced</th>
<th>Article model</th>
<th>Revision model</th>
<th>Quality class model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$: 0.912</td>
<td>$\tau$: 0.896</td>
<td>$\tau$: 0.881</td>
<td>Evenly spaced</td>
</tr>
<tr>
<td>$r$: 0.982</td>
<td>$r$: 0.982</td>
<td>$r$: 0.979</td>
<td>Article model</td>
</tr>
<tr>
<td></td>
<td>$\tau$: 0.984</td>
<td>$\tau$: 0.968</td>
<td>Article model</td>
</tr>
<tr>
<td></td>
<td>$r$: 0.999</td>
<td>$r$: 0.989</td>
<td>Revision model</td>
</tr>
<tr>
<td></td>
<td>$\tau$: 0.983</td>
<td>$\tau$: 0.983</td>
<td>Quality class model</td>
</tr>
<tr>
<td></td>
<td>$r$: 0.989</td>
<td>$r$: 0.989</td>
<td></td>
</tr>
</tbody>
</table>

---

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 measure is not needed. Although it is theoretically possible that using the MPQC might introduce statistical bias because it less accurately captures the measure's ability to model these differences and the space between these classes will grow if its predictive performance improves.

## 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 necessarily 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 ordinal regression model 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 or is not able to improve model prediction because the differences 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 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 of results.

Figure 4: Correlations between quality measures show that the different approaches to measuring quality are quite similar. "Evenly spaced" uses a weighted sum of the ORES scores with handpicked coefficients [16]. Lower values of Kendall’s $W$, a nonparametric rank correlation statistic, compared to Pearson’s $r$ suggest nonlinear differences between the weighted sum and the other measures.

<table>
<thead>
<tr>
<th>Evenly spaced</th>
<th>Article model</th>
<th>Revision model</th>
<th>Quality class model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$: 0.912</td>
<td>$\tau$: 0.896</td>
<td>$\tau$: 0.881</td>
<td>Evenly spaced</td>
</tr>
<tr>
<td>$r$: 0.982</td>
<td>$r$: 0.982</td>
<td>$r$: 0.979</td>
<td>Article model</td>
</tr>
<tr>
<td></td>
<td>$\tau$: 0.984</td>
<td>$\tau$: 0.968</td>
<td>Article model</td>
</tr>
<tr>
<td></td>
<td>$r$: 0.999</td>
<td>$r$: 0.989</td>
<td>Revision model</td>
</tr>
<tr>
<td></td>
<td>$\tau$: 0.983</td>
<td>$\tau$: 0.983</td>
<td>Quality class model</td>
</tr>
<tr>
<td></td>
<td>$r$: 0.989</td>
<td>$r$: 0.989</td>
<td></td>
</tr>
</tbody>
</table>
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 validity 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 greatly vary 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.

7 CODE AND DATA AVAILABILITY
Code, data and instructions for replicating or reusing this analysis are available in the Harvard Dataverse at https://doi.org/10.7910/DVN/US5VGi.

ACKNOWLEDGMENTS
I am grateful to the members of the Community Data Science Collective for their feedback on early drafts of this work including Kaylea Champion, Sneha Narayan, Jeremy Foote, and Benjamin Mako Hill. I would also like to thank Aaron Halfaker for encouraging me to write this after seeing a preliminary version. Thanks to Stuart Yeates and other participants in the wiki-research-1 mailing list wiki-research-l@lists.wikimedia.org for answering my questions about measuring article quality and effort. Finally, thank you to the anonymous OpenSym reviewers whose careful and constructive feedback improved the paper.

REFERENCES