It's All News to Me: The Remix

Miriam Boon
Northwestern University
MiriamBoon2012@u.northwestern.edu

Abstract

This paper examines the automation of editorial curation of online news and blog articles based on reader ratings. Websites usually provide no guidelines on how to evaluate and rate articles; the NewsTrust project explores how doing so could improve rating precision. Building on and expanding from existing, but incomplete, research, I describe simulations of article comparison to determine how many reader ratings are necessary to distinguish between articles.

1. Introduction

Approaches to automating editorial judgment are often based on article ratings by readers. But websites typically give these raters no guidance regarding the basis upon which they should rate content.

The website Slashdot [5, 10, 11, 13] is an exception. It successfully uses user ratings to moderate and curate technology news and comments, identifying funny, insightful, interesting, and informative content. NewsTrust [3, 16] sought to do the same for a broader selection of news and blogs by prompting users with questions related to specific aspects of journalistic quality. Before the project launched in 2005, NewsTrust partnered with academic researchers Cliff Lampe and Kelly Garrett to validate this approach.

Although their study gathered data about both news and blogs, Lampe and Garrett's published analysis [8] did not address the blog data. This paper presents three studies to expand on their results. First, I replicate their study, extending their analysis to the blogs. This series of analyses address several research questions: "How well can these review instruments and questions discriminate between a high- and low-quality article", "How do the rating instruments differ systematically", and "How

Table 1. Questions and survey instruments

Attribute	**Question Wording** (scale of 1-5)	**Full**	**Norm-ative**	**Descr-iptive**	**Mini**
Accuracy	How accurate is this story, overall?	✔	✔		
Credibility	How credible are this story's sources?	✔	✔		
Fairness	How fair is this story, overall?	✔	✔		
Information	How much new information did you get from this story?	✔	✔		
Originality	How original is this story, overall?	✔	✔		
Respect	How respectful is this article towards other viewpoints?	✔		✔	
Balance	How well does this story represent all important viewpoints?	✔		✔	
Clarity	How clear is this story, overall?	✔		✔	
Context	How well does this story help you see the 'big picture?'	✔		✔	
Diversity	How well does the story seek out diverse sources?	✔		✔	
Evidence	How well does this story back its points with factual evidence?	✔		✔	
Objectivity	How well does this story seek out facts, rather than opinions?	✔		✔	
Transparency	How well does this story identify its sources?	✔		✔	
Overall Quality	How do you rate the overall quality of this story?	✔	✔	✔	✔

The distribution of the 14 review questions across the four survey instruments. Note that "Respect" did not make it into the original study, as it was only used for rating blogs.

URI: http://hdl.handle.net/10125/50101
ISBN: 978-0-9981331-1-9

accurate is each review instrument and question with respect to expert journalistic ratings?"

The second study analyzes and compensates for the effect of the sampled population's left-wing bias. Here, my research question has three parts: "Are there indications that the sample's political bias may be affecting the results? If so, can we compensate for that political bias using the existing data? When we do, how do the results change?"

Finally, the third study simulates the comparison of two articles. It is designed to answer the question, "How many ratings are necessary for effective discrimination between articles using the NewsTrust questions and survey instruments?"

2. Related work

Since Garret and Lampe's study, a number of researchers have studied or been inspired by NewsTrust. The NewsTrust Xplorer allowed users to discover and analyze correlations between article keywords, review ratings, and reviewer demographics [2]. NewsTrust data has been used to study the relationship between rater perception of factual content and residual normalized entropy [1]. NewsTrust data has also been used to develop a probabilistic graphical model that leverages the joint interaction of users, news, and sources to make rating predictions [14].

Others have used similar websites, such as Slashdot [4, 6, 9, 10, 12, 13], to: develop an open editing algorithm that determines voter credibility and uses that to in calculating article ratings [7], predicting text trustworthiness using the semantic concept of evidentiality [15], examining the feasibility of using reader ratings to separate high and low quality comments [10] and how users use built-in tools to do such filtering [13], studying the effects of user moderation on new users [9], comparing the effects of user ratings on technology content vs. political content [12], and conducting a statistical analysis of user reaction times to new content [6].

3. The data

NewsTrust constructed four survey instruments, all designed to prime readers to rate stories based on concrete concepts related to journalistic quality. They were called the mini, normative, descriptive, and full reviews with one, six, eight, and 13 questions respectively (Table 1).

They selected one news, and one opinion, article published immediately preceding the experiment, and created a degraded version of each. Then they had three and six expert journalists rate both versions of the news and blog respectively based on all questions appearing on the various review instruments. Experts rating the news stories had a Krippendorff's alpha for interval data of 0.66; the blogs had only 0.425, but selecting the three experts with the most agreement I was able to raise this to 0.64.

The invitation to participate was sent to 6000 individuals who had participated in previous NewsTrust surveys and who had agreed to be contacted again. The original pool was drawn from the membership of MediaChannel.org and MoveOn. This pool suggests that respondents were not representative of the American public, nor online news readers. The 1009 respondents were more liberal, more educated, and older than the average American. They also consumed more news and spent more time on the Internet. Garrett and Lampe felt, however, that "[t]hough the test population is not representative of online news users more generally, we see no reason to expect that their characteristics will influence the relative performance of the instruments. Our finding should hold for other populations."

Each participant was assigned to either the news report or blog, to the original or degraded version, and to one of four review instruments. After reading and reviewing the article, they were asked more questions ranging from basic demographics to queries about interest in news, media use, and journalistic experience, if any (Table A9[1]).

4. Study 1: Replication and extension

Table 2. Discriminatory power, blog

	n	Diff	P > |t|
Full	133	-0.3	0.0500
Descriptive	135	-0.4	0.013
Normative	145	-0.5	0.0004 †
Mini	178	-0.3	0.0415
Average	*591*	*-0.4*	*< 0.0001*

† unequal variances

The original analysis examined the discriminatory power of the four review instruments for the news report, identified systematic differences between instruments, and calculated the accuracy of each instrument with respect to the expert ratings. None of these analyses were conducted for the blog data or the individual questions. In this study I replicate the

[1] http://www.miriamboon.com/newstomeappendix.html

original analysis and extend it to the blog and questions.

The original analysis assumed equal variance. I contrast, I check for unequal variance and use the indicated statistical test. However, even when testing for unequal variance was indicated, it made little or no difference to the outcome.

Where my replication differs from the original analysis, differences are small and likely stem from differences in rounding, truncation, or tests used.

4.1 Discriminatory power of survey instruments and questions

Analysis of the discriminatory power of the review instruments for the news report duplicated the results reported by Garrett and Lampe with a few small differences (Table A1). Generally, results were more significant than they indicated. Also, the rating difference for the full review was -0.5, while they reported it as -0.4.

All of the instruments discriminated significantly between original and degraded news reports. The full review failed to do so for the blog; all others were significant.

For the news, the instrument with the most discriminatory power – and the most error – was the mini followed by the full review. The normative review performed best for the blog, followed closely by the descriptive review (Table 2). Ratings of the original and degraded stories showed no significant difference for the questions connected to "Information," "Clarity," and "Originality" for the news and blog (Table 3). "Transparency" and "Diversity" were not significant for the blog.

The questions with the most discriminatory power for the news were (in order of rating difference) "Transparency," "Context," "Overall Quality," "Evidence," and "Credibility." For the blog, in order of rating difference, they were "Fairness," "Respect," and more distantly, "Overall Quality" and "Accuracy."

4.2 Systematic Differences

Garrett and Lampe did a series of paired t-tests to investigate any systematic differences between the instruments within each condition. I conducted a similar analysis for both the news report and blog.

In almost every case, the mini and normative review resulted in higher ratings; this difference was not always significant, except in the case of the original news report, for which the normative review yielded higher ratings (Table A3). Likewise, the full and detailed reviews almost always yielded ratings lower than the others. Note that this behavior is consistent regardless of article quality and type. Thus, shorter reviews typically yielded higher average ratings.

Table 3. Discriminatory power, questions

		n	Diff	P > |t|
News	Information	205	-0.1	0.4931
	Evidence	175	-0.5	0.0013
	Transparency	178	-0.8	< 0.0001
	Diversity	175	-0.4	0.0171
	Credibility	165	-0.5	0.0012
	Fairness	180	-0.4	0.0079
	Balance	171	-0.4	0.0186
	Objectivity	177	-0.4	0.021
	Accuracy	99	-0.4	0.0237
	Clarity	176	-0.1	0.5722
	Originality	174	-0.2	0.1976 †
	Context	170	-0.6	0.0013
	Overall Quality	417	-0.6	< 0.0001
Blog	Information	275	-0.2	0.1403
	Evidence	266	-0.3	0.0331
	Transparency	263	-0.3	0.1595
	Diversity	242	-0.2	0.0705
	Credibility	233	-0.3	0.0351
	Fairness	253	-1.0	< 0.0001
	Balance	258	-0.4	0.0027
	Respect	240	-0.9	< 0.0001 †
	Accuracy	145	-0.5	0.0045
	Clarity	267	-0.1	0.6107
	Originality	244	-0.1	0.3119
	Context	253	-0.4	0.0046
	Overall Quality	588	-0.5	< 0.0001 †

† unequal variances

There are many possible explanations for this behavior. Participants may hesitate to give low ratings when there are fewer questions, or round up when there are few questions but rate more precisely when there are many, more specific questions. Also, the same question may seem more specific when seen alongside other related questions.

A survey instrument with a lower proportion of problematic questions may perform better. Each

question that is problematic is a potential confound. The process used to degrade the articles did not consistently degrade them along all dimensions. Both the expert ratings and textual analysis of the substance of the changes support this assertion. This may have affected results. Furthermore, three questions may be confounded by reader characteristics.

Table 4. Accuracy, blogs

	n	Abs. Err.	r	P > |t|
Full	133	0.65	0.19	0.027
Descriptive	135	0.65	0.26	0.0027
Normative	145	0.71	0.27	0.0011
Mini	178	0.88	0.15	0.0415

As phrased, both "Originality" and "Information" depend on what the rater has read. To test this supposition, I analyzed the relationship between both questions and news use. As expected, as news use went up, information ratings went down. Their correlation was -0.15 ($p = 0.0007$). Likewise, as news use went up, originality ratings went down, with a correlation of -0.092 ($p = 0.062$).

Table 5. Question accuracy, news

	n	Abs. Err.	r	P > |t|
Information	205	0.81	0.05	0.505
Evidence	175	1.19	0.24	0.001
Transparency	178	1.26	0	1.000
Diversity	175	1.19	0.18	0.017
Credibility	165	0.91	0	1.000
Fairness	180	1.25	0.20	0.009
Balance	171	0.98	0.18	0.018
Objectivity	177	1.06	0.17	0.021
Clarity	176	0.86	0.04	0.570
Originality	174	0.97	0	1.000
Context	170	1.04	0.24	0.001
Overall Quality	417	1.25	0.26	< 0.0001

"Clarity" may depend on rater reading level. However, none of the potentially correlated factors (e.g. news use, education, journalistic experience, web use, or age) had a significant relationship. Measures with strong, significant relationships with clarity also had significant relationships, always in the same direction, with nearly all of the questions. Although experts found small differences between the original and degraded articles for both the news and blog, they may have been mistaken, or referring to issues of clarity to which the general public is not sensitive. Alternatively, there may be a problem with the question as posed: "How clear is this story, overall?" Perhaps not all respondents interpreted the question as having the same meaning.

4.3 Accuracy

Note that "Accuracy" and "Respect" are omitted from this stage of analysis; the former because some experts skipped the question, and the latter because it was not included in the questions experts had to answer.

Nonetheless, results for the news report remained consistent (Table A4): the mini review had the greatest error, and the normative review had the least. The mini, followed by the full review, had the highest correlation with experts. For the blog, the mini review had the greatest error, and the full and descriptive reviews had the least (Table 4). The normative and descriptive reviews had the highest correlation with experts, and the mini review had the lowest.

Respondent ratings for "Information," "Clarity," and "Originality" showed no significant relationship with the expert ratings for news and blogs. For blogs only, respondent ratings for "Diversity" and "Transparency" were likewise insignificant in their relationship with expert ratings.

For news, "Overall Quality," "Evidence", and "Context" had the highest correlations with the experts, and for the blog, "Fairness" had the largest correlation with experts.

Table 6. Question accuracy, blogs

	n	Abs. Err.	r	P > |t|
Information	275	1.02	0.09	0.145
Evidence	266	0.86	0.12	0.0406
Transparency	263	0.87	0.09	0.16
Diversity	242	0.85	0.12	0.0693
Credibility	233	0.98	0.13	0.045
Fairness	253	1.02	0.41	< 0.0001
Balance	258	0.83	0.19	0.0019
Clarity	267	1.06	0.03	0.6091
Originality	244	0.86	0.06	0.3119
Context	253	0.94	0.18	0.0033
Overall Quality	588	0.85	0.22	< 0.0001

4.4 Discussion

No single review instrument was superior in all respects. As Garrett and Lampe suggest, the best instrument may depend on the application. Sometimes, discrimination is important and accuracy

does not matter; other times, priorities are reversed or equal.

The most effective instrument for news and blogs differed, perhaps because the two styles are fundamentally different. For the news, the mini review had the highest discriminatory power and correlation with experts, but also the highest error with respect to experts. The full review offered the best balance between these factors. The normative and descriptive reviews had the best performance for the blog, and the mini review had the worst.

This analysis clearly showed that as phrased, the "Information," "Clarity," and "Originality" questions are of little value, regardless of condition. "Context" and "Overall Quality" proved to have the most value for news, and "Fairness" outperformed all other dimensions for the blog.

That said, while suggestive, I don't think we can draw any firm conclusions based on this data. In addition to confounding issues already discussed, I have only tested these instruments with one news report, and one blog. It remains unclear whether differences I observed are a measure of the article, or the question or instrument being used.

There are some problematic assumptions underlying this analysis. I treat the questions here as being independent of the instrument, averaging them together. The way a respondent interprets a question, however, is dependent on the context, including the other questions on the survey instrument.

I recommend as further work that questions be studied in isolation before being combined into an instrument. Questions should be analyzed for correlation with demographic features to eliminate any with confounds.

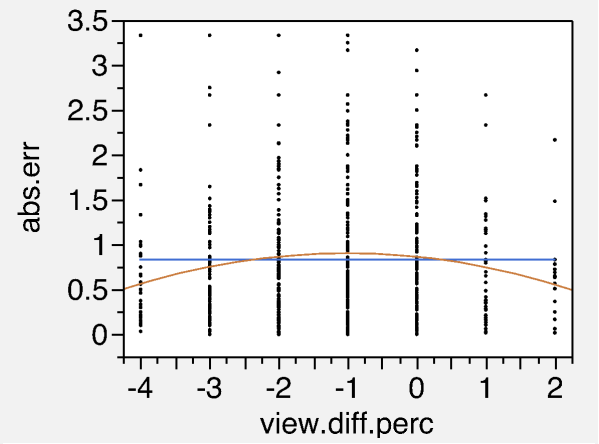

Figure 1. Perceived difference between story and reader's political ideology vs. absolute error.

5. Study 2: Compensating propensities

Study participants were drawn from email lists of political activism groups generally considered left wing. In fact, 82% of respondents identified as "very liberal" or "liberal", while only 2.5% identified as very conservative or conservative. I believe this is a barrier to the generalization of results.

Consider the perceived political difference between raters and the story, calculated by subtracting their rating of the article's political perspective from their own self-reported political stance. A negative perceived difference suggests the participant felt more left-wing than the article, and a positive perceived difference suggested they felt more right-wing than the article.

Plotted against absolute error, this yields a distribution that fits an inverted quadratic with its maximum at -1.2 ($p = 0.0003$, $R^2 = 0.0188$, Figure 1). Thus, the greatest error occurs when readers perceive an article as mirroring their views. A balanced sample of conservatives and liberals would yield a distribution centered on the article's bias; the vertex would shift left for a more conservative article, and right when it is liberal. The maxima for the news (more conservative) and blog (more liberal) ratings (-2.18 and -0.61 respectively) support this supposition.

Table 7. Perceived difference between rater and story, squared, vs. absolute error

	R-Sq	Prob > F
Information	0.026	0.0052
Evidence	0.052	<.0001
Transparency	0.034	0.0017
Diversity	0.018	0.048
Credibility	0.030	0.0062
Fairness	0.070	<.0001
Balance	0.011	0.1336
Objectivity	0.15	<.0001
Accuracy	0.15	<.0001
Clarity	0.042	0.0004
Originality	0.090	<.0001
Context	0.053	<.0001
Overall Quality	0.026	<.0001
Composite	0.019	0.0003

As you can see in Table 7, this curved relationship held for all of the individual questions except "Balance."

In sum, it seems clear that the lack of balance in the study participants is a potential threat to validity. In order to compensate for that, I engaged in a two-step process.

5.1 Identifying unlabeled conservatives

My first step was to supplement the 24 self-identified conservatives by identifying other conservative participants. Respondents who identified as moderates or did not identify their own political affiliation were labeled as unknown. I used a voting features interval (VFI) classifier to classify these participants based on age, news use, web use, gender, income, school, journalistic experience, and the partner organization that referred them to the study. In training this algorithm had a recall of 98.2% and a precision of 98.3%. This process classified nine moderates and one unknown as conservative.

5.2 Modeling propensity

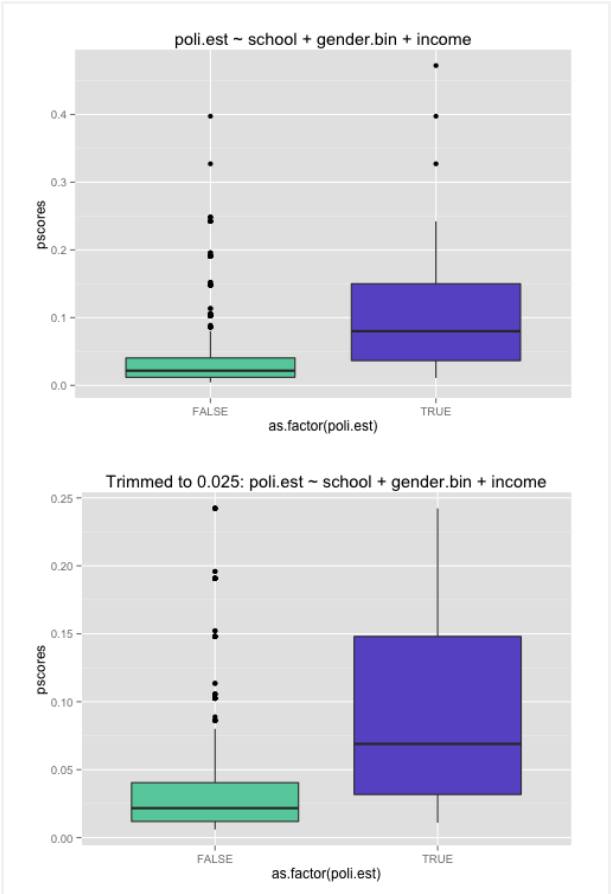

Figure 2. FALSE indicates self-identified liberals. TRUE indicates conservatives, detected and self-identified Top: Distribution of propensity scores. Bottom: Same, with top and bottom 2.5% quantiles trimmed.

In order to model propensity for being conservative (Figure 2), I had to remove participants with unknown political affiliation, and participants with missing data for income, gender, and school. This removed two estimated conservatives, and five of the known conservatives.

To check the efficacy of weighting with the propensity scores, I repeated the analyses from this study's introduction (Figure 3).

As we'd expect if we had added conservative people, the vertex shifted towards zero. Furthermore, the curve has become more pronounced, and the R^2 value has increased, suggesting that the unbalanced sample was obfuscating the importance of this relationship.

Given these indications that the weighting of inferred and self-labeled conservatives manages, to some extent, to compensate for the sample imbalance, I rechecked a number of the analyses to see how they were altered.

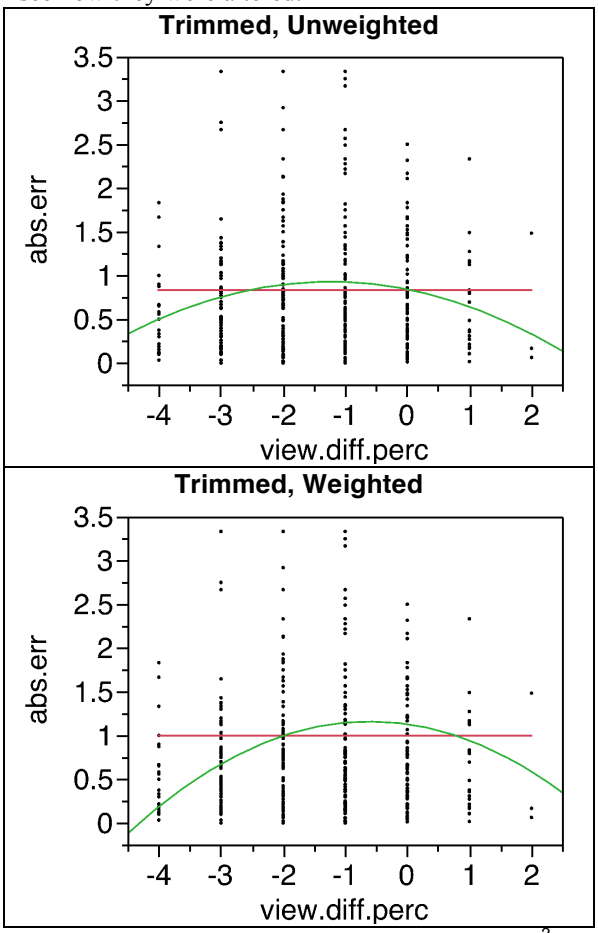

Figure 3. The original vertex was at -1.2, R^2 = 0.019. Top: Trimmed but unweighted data set, centered at -1.4 (R^2 = 0.030, p = 0.0004). Bottom: Trimmed and weighted distribution, centered at -0.8 (R^2 = 0.093, $p < 0.0001$).

5.3 Discriminatory Power

With the original data set, all instruments discriminated between the original and degraded versions of the news and blog, except perhaps for the full review for the blog, and all of the instruments led to scores that would class the original story as higher quality than the degraded version. In all but one case (weighted descriptive instrument for the news report), the effect using the enhanced data set is still clearly pointing in the correct direction (Table 9). However, not all of these differences remain significant. The descriptive review is no longer significant in any of the combinations of conditions, whereas the full and normative reviews are significant for all but the weighted blog (although only to $p < 0.1$ for the unweighted normative news), and the mini review was significant only for the unweighted news and the weighted blog.

As for the size of this effect, with the original data we found that the mini and normative review had the greatest discriminatory power for the news and blog, respectively. This remained true for the trimmed but unweighted data. However, weighted, the normative and mini reviews proved to have the greatest discriminatory power for news and blog respectively.

Table 9. Discriminatory power, trimmed

		Unweighted		Weighted	
		Diff	P > |t|	Diff	P > |t|
News	F	**-0.6**	0.01	**-0.5**	0.0018
	D	-0.4	0.1326	0.3	0.2291
	N	*-0.3*	0.0565	**-0.8**	< 0.0001
	M	**-0.7**	0.0107	-0.3	0.1092
	Avg	*-0.5*	*< 0.0001*	*-0.4*	*0.0001*
Blog	F	**-0.3**	0.0456	-0.3	0.2
	D	-0.3	0.2001	-0.1	0.6323
	N	**-0.5**	0.0185	-0.3	0.1145
	M	-0.2	0.311	**-1.4**	< 0.0001
	Avg	*-0.3*	*0.0008*	*-1.1*	*< 0.0001*

There are a few patterns and changes worth noting with respect to the discriminatory power of the individual questions. As with the original data set, "Clarity" and "Originality" did not show any significant differences. "Evidence" also became insignificant for all combinations of blog, news, weighted, and unweighted, which was not the case with the original data set. "Information," which was insignificant for all configurations of the original data, was significant for the weighted news.

5.4 Accuracy of Ratings

With the original data set, all but the full blog showed a significant ($p < 0.05$; for the full blog $p = 0.0508$) correlation between crowdsourced ratings and expert journalist's ratings. The trimmed data set had fewer significant correlations; the descriptive survey instrument was not significant for the news report or blog, in both the weighted and unweighted calculations. Only three of the weighted correlations had significant values: the Full News, Normative News, and Mini Blog conditions (see Table 13).

Table 10. Accuracy, trimmed data

		Unweighted		Weighted	
		r	p-value	r	p-value
News	F	0.361	0.010	0.430	0.002
	D	0.216	0.133	-0.173	0.229
	N	0.229	0.057	0.547	< 0.0001
	M	0.303	0.011	0.191	0.114
Blog	F	0.243	0.046	0.059	0.632
	D	0.151	0.200	0.151	0.200
	N	0.258	0.019	0.148	0.181
	M	0.097	0.311	0.474	< 0.0001

As before, the mini instrument had the greatest error. For the news report, the least error came from the normative instrument for all data sets (the original, trimmed unweighted, and trimmed weighted). The least error for the blog, for all three data sets, came from the descriptive review instrument.

5.5 Discussion

The differences among data sets representing different balances of political perspective we found argues that there is a relationship between the reliability of a rating and the difference the rater perceives between her own perspectives and the article's stance. This, in turn, suggests that a sample with an overabundance of people with similar political views could be a real threat to validity in studies of this sort.

This method of simulating a more balanced sample is suggestive, but not conclusive. It cannot wholly compensate for a sample drawn from a population that is so drastically skewed. Further research with new, balanced data is required to draw more firm conclusions.

6. Study 3: Simulation

NewsTrust ran from 2005 through 2012, gathering data on a wide variety of articles. Each article on the site was rated at least once, but few had more than three ratings. To rely on this data, we must understand how many ratings two articles must have in order to discriminate between them.

To address this question, I simulated two articles of the same type being evaluated by the same number of respondents, *n*, via the same review tool. My research question was, "How large must *n* be within each condition to discriminate between the original and degraded stories based upon any one of the questions or summary measures?"

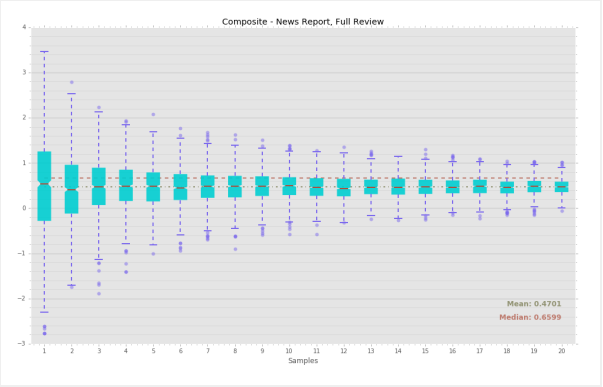

Figure 4. One of 70 graphs generated during the simulation analysis. Number of samples vs. the difference between the original and degraded article, delta, based on the composite scores, from 1000 trials conducted for the news, using the full review instrument.

6.1 Method

Article type (news vs. blog) and tool (mini, normative, descriptive, and full review) gives us a total of eight conditions. Simulation and analysis were conducted using the MiniConda stack, which includes NumPy, SciPy, and Pandas.

Within a condition, I randomly sampled *n* respondents who were shown the original article, where n ranged from 1 to 20. I averaged their ratings for the applicable condition. I then repeated this for *n* participants who were shown the degraded article. This is one trial; for each *n*, for each condition, I conducted 1000 trials.

6.2 Results

I began by visually inspecting the boxplot graphs of the data (Figure 4). As the number of respondents sampled rises from n = 1 to n = 20, the range of values tightens and the median stabilizes at a value greater than zero. This is what we would hope to see.

Next, I conducted two-sample T-tests comparing original and degraded articles for each set of trials. I assumed independence, but did not assume equal variance.

After inspecting the results, I found that in most conditions, the original article was significantly ($p <= 0.05$) rated more highly than the degraded article even *with only one rating each* (Table A10).

The most problematic exceptions were:

- Blog and news articles rated for "Clarity" using the descriptive review instrument did not significantly differ across conditions, and the relationship frequently reversed direction.
- News articles rated for "Information" using the normative review instrument were significant across conditions only for n >= 5, and then the degraded article was actually rated more highly than the original.

In a handful of cases, significance was only achieved for larger values of n:

- Blog, Accuracy, Full review – $n \geq 6$
- Blog, Clarity, Full review – $n \geq 2$
- News, Originality, Normative – $n \geq 3$
- Blog, Transparency, Descriptive – $n \geq 2$

With so many instruments and dimensions proving to be significant and able to discriminate between conditions even for n as small as 1, some other method of determining the most efficacious review instruments was needed. I measured the effect size by calculating delta, the difference between the average ratings of the original and degraded articles. Effect size, however, does not tell the whole story. I propose the concept of *resolution* as another way of looking at these articles.

According to the Oxford English Dictionary, resolution is defined as "The smallest interval measurable by a telescope or other scientific instrument; the resolving power." In this context the resolution is dependent on the value of n. Consider the case, $n = 1$. The rating for each article will always be an integer, and thus delta will be an integer. So when $n = 1$, we do not have the resolution to detect the differences for a pair of articles that differ by only 0.5. However, if at least one article was rated by more than one reviewer, it could have an average rating of 1.5, 2.5, etc. potentially giving a delta of 0.5. So for $n = 2$, we have the resolution to detect differences of 0.5 or more.

I calculated the minimum value of n at which each effect size could have been detected. Only blog articles rated for fairness using the normative review were both significant and resolvable at n = 1. Typical values at which the instruments had the resolution to detect the difference indicated in the simulated trials ranged from n = 2 to 5. The blog rated for accuracy

using the full review had highly variable values for delta, and so it was resolvable for values of *n* ranging from 19-42. Many conditions had smaller variations in delta values that led to a variety of resolution ranges.

- Blog, Clarity, Full review – *n = 5-8*
- News, Clarity, Full review – *n = 5-9*
- Blog, Context, Full review – *n = 5-7*
- Blog, Evidence, Full review – *n = 6-8*
- Blog, Information, Normative – *n = 5-8*
- Blog, Originality, Full review – *n = 9-18*
- Blog, Originality, Normative – *n = 4-8*
- News, Originality, Normative – *n = 9-28*
- Blog, Transpar., Descriptive – *n = 9-21*

6.3 Discussion

These results stem from a combination of article differences and how question sensitivity changes in the context of a given survey instrument. We can explore the latter by asking, "For each rating-article type combination, which instrument has the highest discriminatory power?" Using delta, the rating difference, I determined which instrument consistently yielded the best results (Table A8).

By examining how often a dimension performed particularly well or poorly on a given review instrument, I conclude that the full review is preferable and recommended for news, and the normative is recommended for blogs. Based on these results, I recommend gathering about four samples, which would cover more than half of possible cases. Ten samples would be able to resolve the differences about 90% of the time.

Of course, these recommendations are all based on these specific articles. But we can assume that articles that differ little call for larger values of *n,* and for those that differ a lot, smaller values of n will do.

Note also that *n* was the same for the original and degraded versions of the articles. In practice, two articles will rarely have the same number of reviews. This may impact the resolution calculations.

Also consider that readers often know in advance the current overall article rating. Some may be more likely to rate an article if their opinion contradicts the current rating, others, if it accords. If common, this could represent a threat to this simulation's validity.

7. Conclusion

In order to properly study all the questions raised above, a new experiment must be conducted. However, the present experiment does provide some lessons that will be invaluable to anyone pursuing these sorts of questions. First, the sampled participants must better reflect the general population with respect to political bias. Second, the questions must be tested in isolation before incorporation into an instrument, as some of the questions in these instruments turn out to be highly problematic. As things stand the success of a given instrument had as much to do with the presence of 'dud' questions as it did with the mixture or number of questions included. More than two articles of well-understood, differing quality should be used, and a simulation-style analysis should be included to better understand how the questions and instruments will behave in practice, with varying numbers of ratings.

8. Acknowledgements

I'd like to thank Kelly Garrett and Cliff Lampe for allowing me to use their data. Many thanks also to Darren Gergle, Aaron Shaw, and Larry Birnbaum for their assistance and input. This research was funded by Google.

9. References

[1] Boon, M. 2014. Information density, Heaps' Law, and perception of factiness in news. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. (2014), 33–37.

[2] Cho, H., Kang, C., Oh, U. and Patel, N.S. NewsTrust Xplorer: Content-based Rating Visualization for Exploring News Reviews and Reviewers.

[3] Florin, F. 2009. *NewsTrust Communications 2009 Report*.

[4] Gómez, V., Kaltenbrunner, A. and López, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. *Proceedings of the 17th international conference on World Wide Web* (2008), 645–654.

[5] Halavais, A.M.C. 2001. *The Slashdot effect: analysis of a large-scale public conversation on the World Wide Web*.

[6] Kaltenbrunner, A., Gomez, V. and Lopez, V. 2007. Description and prediction of slashdot activity. *Web Conference, 2007. LA-WEB 2007. Latin American* (2007), 57–66.

[7] Ko, J., Kim, K., Kweon, O., Kim, J., Kim, Y. and Han, S. 2009. Open Editing Algorithm: A Collaborative News Promotion Algorithm Based on Users' Voting History. *Computational Science and Engineering, 2009. CSE'09. International Conference on* (2009), 653–658.

[8] Lampe, C. and Garrett, R.K. 2007. It's All News to Me: The Effect of Instruments on Ratings Provision. *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on* (2007), 180b–180b.

[9] Lampe, C. and Johnston, E. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (2005), 11–20.

[10] Lampe, C. and Resnick, P. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004), 543–550.

[11] Lampe, C., Zube, P., Lee, J., Park, C.H. and Johnston, E. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*. 31, 2 (2014), 317–326. DOI:https://doi.org/10.1016/j.giq.2013.11.005.

[12] Lampe, C., Zube, P., Lee, J., Park, C.H. and Johnston, E. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*. 31, 2 (2014), 317–326. DOI:https://doi.org/10.1016/j.giq.2013.11.005.

[13] Lampe, C.A.C., Johnston, E. and Resnick, P. 2007. Follow the reader: filtering comments on slashdot. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), 1253–1262.

[14] Mukherjee, S. and Weikum, G. 2015. Leveraging joint interactions for credibility analysis in news communities. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (2015), 353–362.

[15] Su, Q., Huang, C.R. and Chen, H.K. 2010. Evidentiality for text trustworthiness detection. *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground* (2010), 10–17.

[16] Survey Report Summary - NewsTrust: 2006. *http://www.newstrust.net/survey/report*. Accessed: 2013-03-21.