It's All News to Me: The Effect of Instruments on Ratings Provision

Cliff Lampe
Michigan State University
409 Comm Arts Building
East Lansing, MI 48130
lampecli@msu.edu

R. Kelly Garrett
University of California, Irvine
3200 Berkeley Place
Irvine, CA 92697
garrettk@uci.edu

Abstract

In this paper, we address an issue of design in online rating systems: how many items should be elicited from the ratings provider. Recommender and reputation systems have traditionally relied on single-dimension ratings to reduce user burden, but for some types of information this amount of feedback may be insufficient. We presented users of an online news rating service with different numbers of items in a news rating exercise. We find that users show the highest satisfaction and greatest rating accuracy with a multi-item reviewing instrument.

1. Introduction

How much information should designers of online rating systems elicit from users in order to balance the burden on feedback providers with the effectiveness of feedback for ratings consumers? Online rating systems collect and aggregate participant feedback on a wide range of goods, including movies, consumer products, comments in online discussion boards, and sellers in online auctions. Feedback from users who have experience with the item being rated is provided to potential future users in order to help them decide how to allocate scarce resources such as attention and money. Ratings based on a single item provide a low-burden approach to collecting this information, but multiple measures may more completely capture the rater's experience.

Most online recommender systems have adopted a single overall-quality measure, such as the 5-star rating used in Amazon's book rating system (http://www.amazon.com/). The ratings are Likert-type evaluations of the user's satisfaction with the good, interaction or service, which are then typically aggregated to provide an overall rating. In some cases, these averages are reported as overall recommendations (e.g., Amazon). In other cases, the recommendations are personalized by algorithmically matching users' preferences. Some systems, like those employed by news discussion site Slashdot (http://slashdot.org), use "thumbs up/down" styles of rating that adjust the aggregated evaluations, rather than averaging absolute ratings. In either case, users are asked to provide only one datum: their overall evaluation of the item being rated.

NewsTrust (http://www.newstrust.net/) is a non-profit organization with the goal of creating an online news rating service to help non-expert reviewers rate news stories in terms of their journalistic quality, rather than their personal appeal. NewsTrust's earliest prototypes of their news rating tool required raters to answer 19 questions during the rating process, but workload and response rate concerns led the group to seek other, less burdensome assessment strategies. The key design objective inspired by this specific need is to determine which assessment strategies provide the optimal balance between rater burden and assessment accuracy.

1.1 Rating online news

An estimated 39.3 million people read online news stories in October 2005, not including those who read blogs [1]. According to a recent report, more broadband users get their news from online sources than they do from their local newspaper [2]. Today, most online news readers rely on sources produced by the major news organizations, but use of alternative sources is on the rise [3]. News rating systems that provide feedback about the quality of these stories and sources could be a valuable resource for readers in choosing content to which to allocate their time.

Designing a rating system for online news poses several challenges because news perceptions are uniquely susceptible to the influence of prior attitudes, and evaluations of news credibility and quality are known to be influenced by raters' opinions about the events being reported [4-6]. For example, people quickly accept evidence that supports their beliefs, but tend to be more critical of belief-challenging

Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07)
0-7695-2755-8/07 $20.00 © 2007 IEEE
1530-1605/07 $20.00 © 2007 IEEE

information. These biases make it particularly difficult to capture "objective" news ratings. One way of dealing with these biases is to collect ratings on a range of story characteristics. Collecting more contextual information may be better than simple statements of preference because it provides information about the characteristics that influence raters' quality assessments.

1.2 Online rating systems

Rating systems are useful when the past is predictive of the future [7], such as when one person's enjoyment of an experience indicates that a future user might enjoy it as well. Rating systems help provide information about uncertain choices by using previous users' experiences to make recommendations to future users.

Researchers have identified a variety of factors that influence the accuracy of rating systems. In reporting on the development of GroupLens, Miller et al. [8] tested correlations between predicted quality scores and actual user ratings, and found that the strength of the relationship (that is, the accuracy of the prediction) depended on the nature of the content being rated. For example, on Usenet humor posts seemed to have clearer, more consistent guidelines within the user community about what made a "good" post than recipe posts. As a result, software-generated predictions regarding humor posts were more accurate than predictions about recipes. Herlocker et al. [9] expand the methods used to evaluate the accuracy of ratings to include dimensions of novelty, serendipity, confidence in predictions, coverage of ratings, and dimensions of user evaluation. Lampe and Resnick looked at dimensions like agreement between raters and evaluations of the fairness of ratings to determine the accuracy of ratings in the Slashdot comment system [10]. They found that users of that site did broadly agree on what constituted a "good" comment, with little disagreement between raters on the feedback a comment should receive. In other work, however, Lampe and Johnston [11] found that ratings differed significantly in threads that were marked as technology-oriented versus those that were politics-oriented. In political threads, comments received more ratings and the ratings were more contentious. Provision of ratings is affected by the content being rated.

In the literature on online rating systems, there has been little discussion of the appropriate number of dimensions on which to rate an item. Konstan et al. [12] describe a design specification for GroupLens as a single-dimension rating, since "users typically spend very little time or attention on any particular article," though the basis of this assertion is not reported. In discussing the evaluation of recommender systems, Herlocker et al. [9] mention that most recommender systems, especially those for consumer goods, have used single-dimension ratings. Miller et al. [8] assert that users would prefer not to rate at all, and suggest that recommender-systems researchers seek ways to further reduce the burden of rating. Specifically, they call for research comparing explicit, user-provided ratings with implicit measures, such as the number of times content is accessed. Avery et al. [13] have described the provision of feedback in terms of a public goods problem. Users can "free ride" on the work of those who initially provide ratings without providing ratings of their own. This creates an incentive problem in generating the ratings in the first place. No empirical work compares the value of single-dimension rating versus multiple dimension rating in recommender systems. The design assumption that less user burden is *always* better seems to derive from more general principles of usability design.

Researchers have examined other ways of adding rating context, though. Many online rating systems supplement numerical data with detailed written reviews of users' experiences. eBay users are encouraged to leave written feedback about transactions as well as quantitative rating of the experience. This technique can provide compelling data to future consumers of the rating, but compounds the workload problem faced by the initial providers of the feedback.

While these methods of adding explanatory power to ratings are interesting, little attention has been paid to how multi-item scales might also add nuance to rating systems.

1.3 Rating systems as surveys

A common justification for the use of single-dimension rating in online ratings systems is to increase participation by reducing participant workload. Survey research methodology has examined how the number of items in a survey influence participation rates, which they frame in terms of non-response bias [14]. Bogen [15] reviews decades of work on the effect of questionnaire length on response rates, and concludes that there is no clear relationship between the number of items in a survey and how many people will participate in it. Dillman [16] studied several different versions of the U.S. Census looking at the effects of instrument length, and found only marginal differences between the shortest and

longest versions of the Census instrument in terms of participant drop out rates.

Like designers of recommender systems, the basic premise of survey questionnaire designers has been that short instruments reduce the burden on respondents. Converse and Presser [17] refer to this as reduced cognitive burden, though they do not point to any empirical evidence on the effects on respondents. Unlike designers of recommender systems, the creators of survey instruments have studied the use of multiple item scales in eliciting preferences and attitudes. Multiple questions on a single topic provide context to complex subjective evaluations [17]. Although survey designers mention the need to balance the length of the instrument with the benefits of multiple questions on a single topic, the general design principle is that more items elicit more accurate viewpoints from respondents [18].

Research has also shown that questions focusing on specific attitudes or behaviors generally elicit more accurate feedback than those framed in terms of abstract concepts [17]. Responses to broader, more general questions are more likely to be swayed by top-of-the-head considerations [19]. This suggests that detailed, descriptive questions about specific attributes of the content being rated could yield more consistent ratings. On the other hand, if the descriptive questions are too difficult to answer, for example if they require expertise that subjects lack, subjects may revert to guessing, resulting in inconsistent and inaccurate results.

The literature from survey research indicates that multiple-item scales are preferable in a number of situations. However, both the free-riding problem and usability principles indicate it is preferable to reduce the burden we place on users. This paper addresses that tension: what is the optimal balance between the nuance provided by multiple measures in rating content and the burden of eliciting data from users?

2. Methods

In order to compare the performance of several alternatives to the single-dimension ratings prevalent in recommender systems today, we designed a study of users of NewsTrust, an online news rating service in the early stages of development. This section describes NewsTrust, and the data collection process.

2.1 Newstrust

NewsTrust is a non-profit organization that is developing a rating instrument intended to allow non-experts to provide unbiased assessments of news stories' journalistic quality. The project was founded by Fabrice Florin in 2005 and is scheduled for full public release in 2006. NewsTrust is interested in encouraging raters to evaluate news items based on "journalistic quality" rather than popularity or ideology. To that end, the organization initially developed a questionnaire based on editorial guidelines used by major news organizations. The questionnaire included items related to accuracy, fairness and originality. These dimensions were generated from a

Table 1: NewsTrust rating instrument attributes and question wording.

		Large	Medium		Small
Attribute	**Question wording** (scale of 1-5)	Full Review	Normative Review	Descriptive Review	Mini Review
Accuracy	How accurate is this story?	✓	✓		
Credibility	How credible are this story's sources?	✓	✓		
Fairness	How fair is this story?	✓	✓		
Informativeness	How much new information did you get from this story?	✓	✓		
Originality	How original is this story?	✓	✓		
Balance	How well does this story represent all important viewpoints?	✓		✓	
Clarity	How clear is this story?	✓		✓	
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 it support 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?	✓	✓	✓	✓

review of the codes of ethics and editorial guidelines of several major journalistic organizations, including the BBC, New York Times, International Federation of Journalists and the Washington Post, among others.

2.2 Rating instrument and evaluated content

In order to evaluate the influence of assessment strategy on performance, the research team created four rating instruments. One instrument used a single-dimension rating, and the other three used multiple items. The instruments were composed of five-point Likert-scaled assessments questions, with responses coded such that higher values corresponded to more positive evaluations of content. For the multi-item instruments, the rating is the average of all included items. The instruments differed in two regards. They included different numbers of ratings questions, and they included different types of questions. The specific composition of the review instruments is described below, and summarized (with question wording) in Table 1.

The single-dimension rating instrument, which we refer to as the *mini review*, is similar to those used by most rating services. It asks respondents, "How do you rate the overall quality of this story?" with responses on a five-point scale from very bad to very good. The second instrument, the *normative review*, adds five additional questions. These questions were based on core principles of good journalism, which were derived from the codes of ethics of several news organizations as described above. Subjects were asked to characterize the overall quality of the story in terms of how credible the sources were, how fair the story was, how accurate the story was, and how original the story was. The third instrument, the *descriptive review*, also includes the overall quality measure, but adds several different questions about the core journalistic principles. These questions generally focus on more narrowly defined characteristics of the story. For example, instead of asking how "fair" a story was, this instrument asked if the story includes diverse sources and represents all relevant viewpoints. Descriptive questions are intended to elicit evaluations of specific attributes of the story, in contrast to the summative appraisal elicited by the normative review. The fourth instrument, the *full review*, combined the questions from the other three rating systems. This assessment instrument included 13 items, representing both normative and descriptive style questions.

The instruments were used to evaluate two different versions of the same news article. A panel of professional journalists working with NewsTrust selected a news story published just days prior to the start of the experiment. We refer to this as the *original* story. The group then created a reduced-quality version of the story by introducing a variety of problems, including errors and unsupported opinions, which we term the *degraded* story. The result was a pair of stories that were comparable except in terms of their quality.

2.3 Online experiment

The experiment was administered over the web. To participate in the study, subjects used their own computer, Internet access, and web browser to access a URL provided in the email invitation. From that website, users interacted with a software application that managed the presentation of the news story and assessment instrument, and recorded their responses. Participants were assigned one of two stories to review (original or degraded), using one of four review instruments (mini, normative, descriptive, and full), selected at random. We did not tell participants about our story quality assumptions.

We sent invitations to participate in the web-administered experiment via email to about 6,000 individuals on December 15th, 2005, and subjects had up to one week to complete the study. The email source and return address was associated with the NewsTrust domain. The email invitation for respondents assigned to a full review informed them that the survey would take about 20 minutes to complete; other respondents were told the survey would take 15 minutes for the detailed and abstract reviews and 10 minutes for the mini review.

A total of 418 people responded to the invitation and completed the survey (7% response rate). Invitations were sent to individuals who had previously participated in NewsTrust surveys conducted between March and May 2005, and who had agreed to be recontacted. These respondents were originally recruited from the membership of MoveOn.org and MediaChannel.org. As a consequence, they are more liberal than most Americans: three in four respondents (75%) identified themselves as being politically liberal. We also suspect that respondents are unusually politically interested. Higher political interest is often correlated with higher political sophistication, and so these subjects may tend to perform better than a more typical user.

Respondents were also unique in terms of other demographic characteristics. Older Americans were disproportionately represented: two in five (44%) were between the ages of 50 and 55, compared to only one in five (21%) between 35 and 49, and one in ten (10%) between 25 and 34. Respondents were also

exceptionally well-educated, with more four in five (83%) holding a college degree. Men and women were about equally represented in the sample (51% male).

Though 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.

3. Results

The results of this study are divided into four sections. First, we compare the review instruments in terms of their ability to discriminate between high and low quality news. Next, we compare the ratings generated by the four instruments. In the third section, we evaluate the accuracy of the ratings associated with each review instrument. The final section addresses the question of user burden.

3.1 Users can discriminate between story versions

Table 2 reports the average ratings of news stories in the original and degraded news story conditions by assessment instrument. The table also shows the standard deviation of ratings in each condition, and a t-test reflecting the significance of the difference in average scores between the high- and low-quality conditions. The results show that raters were able to accurately discriminate between original and degraded content using all four rating systems, though the mini review had the most discriminatory power.

We suggest that the reason the gap between the original and degraded story conditions was so much bigger with the mini review than with the other instruments is that raters are more confident in their ability to provide a general assessment of a story than their ability to assess it in terms of its specific attributes. Stating an overall impression is similar to stating an opinion, and people have many opportunities to practice forming opinions about the news. The mini review allows raters to interpret "quality" in their own terms, enhancing their confidence in the legitimacy of their opinion and encouraging them to express more extreme views. Asked to rate content in more specific, but less familiar terms, raters may feel less certain of their opinions and therefore more likely to provide ratings that tend toward the middle.

Table 2. Average quality ratings by news type and review instrument

		Original Rating		Degraded Rating		Rating	
	(n)	Mean	(SD)	Mean	(SD)	Diff.	t-test
Full (13Qs)	(81)	3.0	(.82)	2.6	(.77)	0.4	2.64*
Descript (8Qs)	(97)	3.0	(.90)	2.6	(.88)	0.4	2.34*
Norm (6Qs)	(126)	3.3	(.69)	3.0	(.82)	0.3	2.48*
Mini (1Q)	(114)	3.7	(1.06)	2.9	(1.24)	0.8	3.65**
News Average	*(418)*	*3.3*	*(.92)*	*2.8*	*(.96)*	*0.5*	*5.49***

Note: * p< .05 ** p< .01 *** p < .001

The smallest standard deviations are associated with the normative review instrument. As a consequence, this instrument had a larger t-score than the descriptive instrument even though the difference between high and low quality ratings was slightly smaller. This suggests that the summative questions unique to this instrument were clearer and more easily understood than the more detail-oriented questions found in the other two multi-item instruments, and that they tended to produce more consistent assessments.

One final observation based on a visual inspection of this data is that the longer instruments tend to elicit lower ratings. We examine this trend more closely in the next section.

3.2 Detailed questions lead to lower scores

The previous section showed that the ability to distinguish between an original news story and a degraded version was affected by the type of rating instrument used. Turning to the average scores of the original news stories, we observe that the different instruments are associated with systematic rating differences. Table 3 compares the average ratings across the four review instruments when assessing the original news items. Mini-review scores were significantly higher than scores generated by all other assessment instruments. Additionally, scores achieved through the normative review were marginally higher than those from the descriptive instrument condition.

Instrument-based differences were also evident in the rating for the degraded news story, though there

Table 3. Rating Differences – Original News Story

Review		Full Review		Descriptive Review		Normative Review	
	Avg	Diff	t	Diff	t	Diff	t
F	3.03						
D	3.00	-.03	0.18				
N	3.30	.27	1.66	.30	1.95†		
M	3.65	.62	2.92**	.66	3.41**	.35	2.07*

Note: † p<.1 * p< .05 ** p< .01 *** p < .001
Letters denote instrument F-ull, D-escriptive, N-ormative, or M-ini

Table 4. Rating Differences – Degraded News Story

Review		Full Review		Descriptive Review		Normative Review	
	Avg	Diff	t	Diff	t	Diff	t
F	2.56						
D	2.57	.01	.07				
N	2.96	.40	2.64**	.39	2.43**		
M	2.86	.30	1.53	.29	1.35	-.10	.51

Note: * p< .05 ** p< .01 *** p < .001
Letters denote instrument F-ull, D-escriptive, N-ormative, or M-ini

were fewer of them. Table 4 compares the average ratings between the different review instrument conditions for this content. Scores in the normative review condition were higher than in either the descriptive or full review condition.

Compared to the normative review, the full and descriptive review both generated ratings that were significantly lower for both the original story and the degraded version of that story. *But do the lower ratings more accurately reflect the content being rated?* We address this question in the next section. We begin by describing how we measure accuracy, then we examine the influence of the various rating instruments.

3.3 Instrument type affects rating accuracy

As we noted in our review of the literature, there is little research on the effect of instrument length in eliciting ratings. The assumption has been that brief questionnaires that minimize user burden are at least adequately accurate. We test this claim explicitly by examining the influence of instrument type on rating accuracy. As a reminder, the different instruments had different numbers of questions: the mini review asked only one question, the normative review asked a short series of general questions, the descriptive review asked a short series of more specific questions, and the full review included both the normative and descriptive questions.

Accuracy, in this study, is defined as agreement with expert raters. This definition stems from NewsTrust's goal of helping novice raters evaluate the quality of articles in a manner similar to professional journalists. The benchmark scores in this study are based on the independent ratings of three experienced journalists associated with NewsTrust, and were computed separately for each instrument. The benchmark for the full instrument is the average of all items across the three judges. The benchmark for the smaller rating instruments are averages based on the corresponding subset of questions. The resultant benchmark scores are shown in Table 5. There was an adequate level of agreement among the experts regarding story ratings, with an absolute inter-class correlation coefficient across the three raters of .65.

Table 5. Expert reviewers rating benchmarks by instrument

	Story type	
	Original	**Degraded**
Full	2.57	2.14
Descriptive	2.25	1.75
Normative	3.07	2.70
Mini	2.33	1.67

To test instrument rating accuracy, we borrow methods pioneered by Miller et al [8] and used by others [9]. Miller and colleagues describe four measures of rating efficacy: the mean absolute error (*Err*), the mean squared error (Err^2), the standard deviation (σ) of the error, and the correlation (r) between subject and expert ratings. Mean absolute error represents the average difference between subject and expert ratings. Lower values indicate better performance. The mean squared error penalizes large errors. By this metric, an instrument with consistently moderate errors will perform better than one that generates an even mix of high and low scores. The standard deviation of the error reflects the range of errors. A low standard deviation, when paired with low mean errors, is optimal.

Table 6. Accuracy of ratings by instrument and story quality

Instrument	$\overline{Err^2}$	$\overline{\lvert Err \rvert}$	σ	r
Full	.80	.69	.06	.285**
Descriptive	1.39	.94	.07	.234*
Normative	.64	.62	.05	.218*
Mini	2.90	1.43	.09	.326**

Note: * p< .05 ** p< .01

As shown in Table 6, the full and normative reviews were most accurate. The mean absolute error and mean square error for these instruments were much lower than for either the mini or the descriptive review. It is noteworthy that the mini-instrument condition, which is typical of most recommender systems, introduced the largest error in the ratings process. As a result, the mini review tool is considered least effective for the purpose of eliciting expert-like ratings. In terms of correlation between subject and expert ratings, however, the mini review performed best, followed by the full review. Thus, it appears that the mini review tool is more effective at capturing relative changes than at generating an accurate absolute score.

On the question of accuracy, choosing between the normative and the full reviews requires selecting which metrics to prioritize. The full review presents a unique mix, offering the second highest correlation and the second lowest error rates. If those two factors are considered equally important, then these data suggest that the full review is best. The normative review, however, had the lowest error rate, and still generated a significant positive correlation between subject and expert rating. Thus, if error is the more significant factor, the normative review should be preferred.

3.4 Instrument type affects user burden

Another way to assess the influence of instrument length and question type on news rating is to consider the burden of each instrument on reviewers. Though the recommender system and survey literatures disagree regarding the value of multiple measures, they both hold to the design principle that opinion-seeking instruments should not cost the participant undue time or energy. To determine the effects of the instruments on participant "cost" we look at three measures of user burden: completion rate, completion time, and perceived effectiveness of the rating instruments.

3.4.1 Response rate as a measure of burden

Table 7 reports the response rates for the four rating instruments. The full review, which included 13 rating questions and several questions about the user and the instrument, yielded the lowest response rate. Significantly fewer participants completed the study using these instruments than with either the descriptive or the mini reviews. Though the longest instrument had the lowest response rate, the relationship between response rate and the number of questions asked is not strictly linear. The descriptive review actually had a slightly higher response rate than the mini review. It could be that participants were motivated to complete the short instrument because it seemed more useful or credible (evidence for this explanation is presented below). In sum, these findings show that the simple narrative of "shorter is better" does not stand to scrutiny in this case. There are factors other than questionnaire size at work.

Table 7. Completion rate by instrument

Instrument	Invited	Complete	Response rate
Full	1543	81	5.25%
Descriptive	1543	97	6.29%
Normative	1543	126	8.17% ** [a]
Mini	1543	114	7.39% * [a]
Overall	*6174*	*418*	*6.77%*

Note: * p< .05 ** p< .01 *** p < .001
a. Compared to full review response rate

3.4.2 Time to completion as a measure of burden

Another way to measure the effect that different instruments had on user burden is to measure the time it took respondents in different conditions to complete their evaluations. Table 8 shows the number of minutes users in each condition took to complete the evaluation. Completion time is measured from first site access to questionnaire completion. It includes time spent reading and rating the news story, and providing demographic information.

As this experiment was administered online, we had little control over when participants completed the study. They could leave the experiment web site at any time and return later to finish their evaluation. A number of subjects did this, and the completion times for many of these individuals fell more than 1.5 times the IQR above the third quartile of response times. We assume that these outlier cases, representing subjects who spent more than 41.5 minutes reading 828-word news story (plus title/byline) and answering between 12 and 24 questions, are the result of factors external to the study, and we exclude them from this analysis. The average time to completion for excluded subjects was 12 hours and 14 minutes, dwarfing the three-quarters of an hour threshold.

Although review length influenced completion time, the difference between the longest (full) and the shortest (mini) instrument was less than four minutes on average. The full review had 11 more questions than the mini review, meaning that respondents spent about an additional 20 seconds per question. This time difference suggests that the largest portion of the review process is spent not in assigning evaluations, but reading and considering the content.

Table 8. Completion time (in minutes) by instrument

Instrument	(n)	Mean time to completion	Std. Deviation
Full	(74)	17.03 [a]	9.69
Descriptive	(91)	15.48	7.83
Normative	(116)	15.78	7.72
Mini Review	(92)	13.28[a]	6.29
Overall	(373) [b]	15.34	7.94

Note: a. Difference significant at p<.05
b. 45 outliers, with completion times greater than 41.5 minutes were excluded.

Although it is obvious that more questions take more time, the interesting finding here is that the amount of time between conditions was not all that different. These findings may be different for tasks

where no consideration is necessary, for example a product review where all consideration has happened prior to the rating process. However, these findings may apply very well to contributions to online communities, where the rating happens concurrently with the experience of that which is being rated.

3.4.3 Perceived effectiveness as a measure of burden

Another way the length of the instrument may affect user burden is by affecting how users perceive the effectiveness of the instrument. This was measured by the question, "How well did this review tool help you evaluate the quality of the story?" Responses were given on a 5-point Likert scale anchored by "not well at all", scored as 1, and "very well", scored as 5. Table 9 reports mean responses by condition. Although the normative and mini reviews generated the largest response rates, and the mini review typically took less time to complete, both were perceived by subjects to be less effective than the descriptive review. This is in stark contrast to our accuracy metrics, which show that the normative and full reviews produced results that were generally more in line with experts. We suspect that people are responding to the difficulty of using the rating instrument: because the questions are hard to answer, subjects feel that the instrument promotes more careful scrutiny.

Alternatively, the disparity between perception of efficacy and our objective measures of accuracy might result from the motivation for rating. Users might be hoping for more active involvement in the evaluation process, and consequently rate the conditions in which they have less individual say as less effective. It would be interesting in the future to also measure willingness to rate again based on the use of the different instruments.

Table 9. Perceived effectiveness by instrument

Instrument	(n)	Mean perceived effective	Std. Deviation
Full	(79)	3.48	.92
Descriptive	(94)	3.72 [a, b]	.84
Normative	(118)	3.19 [a]	1.07
Mini	(107)	3.10 [b]	1.03
Overall	(398)	3.35	1.00

Note: a, b. Difference between instrument means with same subscript are significant at p<.05

4. Discussion

Online rating systems play an important role in reducing information overload, shaping user behavior, and predicting user preferences. However, as online rating system use generalizes to endeavors beyond the evaluation of consumer goods or personal preferences, new questions arise. Can all content be rated in the same way? The number of questions is an important design decision that might depend on the type of information being rated.

In this study, we address a concrete design question about the number of items in a rating instrument in order to determine what format is most effective for evaluating online new items. We conducted an online experiment where more than 400 subjects rated news stories using one of four rating instruments. To determine the effectiveness of the different instruments, we considered accuracy and user burden. We looked at a diverse range of factors including error rates, correlations between users and experts, response rates, task complete time, and user-perceived effectiveness.

Table 10. Comparing relative performance of instruments

	Full	Norm-ative	Descrip-tive	Mini
Discrimination	Mid	Mid	Worst	Best
Accuracy	Mid	Best	Mid	Worst
Completion Rate	Worst	Best	Mid	Mid
Completion Time	Worst	Mid	Mid	Best
Perceived Effectiveness	Mid	Mid	Best	Worst

Table 10 summarizes the outcomes for the different rating instruments. No instrument was consistently best on all measures of rating success, but the success patterns provides design guidance for deciding when different types of instruments are best implemented. These data indicate that the goals of rating should strongly influence which type of instrument is most appropriate to use.

When discriminating between choices, the mini tool, common in recommender systems, performs the best. Based on prior research and current practices, we anticipated that the single-dimension rating would be the most effective for all purposes, as it reduces user burden and should, over large numbers of users, provide an accurate assessment. The prevalence of this approach to collecting rating in recommender systems seemed to argue for a prima facie acceptance of the method. In this test, the single-dimension rating was effective for discriminating between the average and low quality conditions, though the ratings tended to

differ substantially from those given by experts. This suggests that the technique is most useful when trying to discriminate between high and low quality choices, rather than provide information about the absolute quality of those choices. We do not assert that single-item rating systems are useless or invalid, but their poor performance in this test raises interesting questions about how instruments effect rating provisioning.

When trying to match the ratings of experts, we find that a multi-item review instrument that included six normative questions performed best. Of the four alternatives, this instrument provides the best balance between user burden and instrument accuracy. It had the lowest error and the highest response rate.

This was not a clear case of more questions being better. By almost every measure, the descriptive instrument performed the worst. This could be the result of conservatism in rating caused by reviewer uncertainty. When a reviewer is not sure how to judge a particular dimension of quality, he or she may gravitate towards the middle. This suggests that when asking non-experts to provide ratings, it is important to be careful not to ask question that are too detailed, too hard to assess, or that are likely to trigger insecurity about their decision.

Given the relative differences in the success of the rating instruments, we assert that different types of content may require different types of ratings. In this case where agreement with experts was the desired outcome, the single-dimension instrument did not fare as well a multiple dimension instrument. However, the single-dimension instrument did take less time absolutely, so in cases where speed is the most important factor may be the preferred choice. Other genres of content where the normative tool may be more effective include online conversations, rating interactions with other users, rating longer pieces of content, or any other rating condition where it would be useful to do more than discriminate between good and bad content.

5. Limitations

There are several limitations of the study that should be noted. First, the analysis of ratings reported in this paper was based on single story in original and degraded form. As a result, we were unable to represent the full range of quality (and flaws) evident in the news media today. For example, some stories are bad in dramatic and obvious ways; others' flaws are more subtle. It may be that longer instruments were better at capturing the types of errors represented by the stories we selected. A useful remedy would be a follow up study examining rating accuracy across a wide range of real news stories over an extended period of time.

Second, while this was in some ways a test of the single-dimension rating common in recommender systems, the participants in the mini review tool item were asked other questions. Collecting demographic data and other measures could affect some of the analyses above. Also, the question we posed for the mini review focused on "the overall quality of this story". Different wording that measured other aspects of the user's experience with the story might have produced different results. Possible follow-up work might address this by including more types of single input responses for comparison.

Third, we have reason to believe that the subjects in this experiment were unusually politically interested and disproportionately represented the political left. Though we do not believe that this affected the relative performance of the instruments, it would be useful to verify this.

Fourth, the method of participation may have elicited different behaviors than would typical interaction with a rating system. By soliciting people to participate in a study, we might have created an atypical motivation to rate, which might affect the findings presented above. A way to overcome this limitation is to devise future studies in which participants rate content without being initially aware they are participating in a study.

Finally, while we believe that the rating of news is a valuable and important endeavor, we are unsure of the extent to which the lessons learned here generalize to other types of content. For example, it may be that when rating consumer goods, multiple items do not improve performance, or even aversely influence response rate.

6. Conclusions

Increasingly, designers are turning to online rating systems to help make sense of an ever-growing universe of information. We conclude from our work that some rating tasks should break from the unitary measures traditionally found in recommender systems. When the topic is complex and the rating is intended to correspond to an institutional norm, not just personal proclivities, more complex rating systems may be appropriate. We have demonstrated that it is possible to use a multi-item rating instrument to improve rating quality without sacrificing user satisfaction. However, we also found that more items are not necessarily better. Highly detailed descriptive question typically generate less accurate results than any other type. As

the prevalence of online rating systems grows, researchers should continue to explore the different contexts of provision and use of ratings.

7. Acknowledgments

This paper benefited greatly from the feedback of the anonymous HICSS reviewers. We would also like to express our gratitude to Fabrice Florin and his collaborators at NewsTrust for their assistance with the experiment. The research would not have been possible otherwise. Thanks to Howard Rheingold for introducing us to this valuable project. Finally, thanks to Ming Liu for her assistance with data analysis.

8. References

[1] T. Yen and S. Bausch, "Online Newspapers Enjoy Double-Digit Year-Over-Year Growth," Nielsen//NetRatings November 15, 2005 2005.

[2] J. Horrigan, "Online News: For many home broadband users, the internet is a primary news source," Pew Internet & American Life Project, Washinton D.C. March 22, 2006 2006.

[3] J. Horrigan, K. Garrett, and P. Resnick, "The Internet and Democratic Debate," Pew Internet & American Life Project, Washington, D.C. October 27, 2004 2004.

[4] C. G. Lord, L. Ross, and M. R. Lepper, "Biased assimilation and attitude polarization: the effects of prior theories on subsequently considerd evidence.," *Journal of Personality and Social Psychology*, vol. 37, pp. 2098-2109, 1979.

[5] G. D. Munro, P. H. Ditto, L. K. Lockhart, A. Fagerlin, M. Gready, and E. Peterson, "Biased Assimilation of Sociopolitical Arguments: Evaluating the 1996 U.S. Presidential Debate," *Basic and Applied Social Psychology*, vol. 24, pp. 15-26, 2002.

[6] R. P. Vallone, L. Ross, and M. R. Lepper, "The hostile media phenomenon: biased perception and perception of media bias in coverage of the Beirut Massacre," *Journal of Personality and Social Psychology*, vol. 49, pp. 577-585, 1985.

[7] P. Resnick and H. Varian, "Special Issue on Recommender Systems," *Communications of the ACM*, vol. 40, 1997.

[8] B. N. Miller, J. T. Riedl, and J. A. Konstan, "Experiences with GroupLens: Making Usenet Useful Again," *Proceedings of the 1997 Usenix Winter Technical Conference*, 1997.

[9] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, pp. 5-53, 2004.

[10] C. Lampe and P. Resnick, "Slash(dot) and burn: distributed moderation in a large online conversation space," presented at Conference on Human Factors in Computing Systems (CHI), Vienna, Austria, 2004.

[11] C. Lampe and E. Johnston, "Follow the (Slash) dot: Effects of Feedback on New Members in an Online Community," presented at International Conference on Supporting Group Work, GROUP '05, Sanibel Island, FL, 2005.

[12] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Applying Collaborative Filtering to Usenet News," *Communications of the ACM*, vol. 40, pp. 77-87, 1997.

[13] C. Avery, P. Resnick, and R. Zeckhauser, "The Market for Evaluations," *American Economic Review*, vol. 89, pp. 564-584, 1999.

[14] R. M. Groves, R. B. Cialdini, and M. P. Couper, "Understanding the Decision to Participate in a Survey," *The Public Opinion Quarterly*, vol. 56, pp. 475-495, 1992.

[15] K. Bogen, "The Effect of Questionnaire Length on Response Rates:A Review of the Literature," presented at American Statistical Association Section on Survey Research Methods, Alexandria, VA, 1996.

[16] D. A. Dillman, M. D. Sinclair, and J. R. Clark, "Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys," *Public Opinion Quarterly*, vol. 57, pp. 289-304, 1993.

[17] J. M. Converse and S. Presser, "Survey Questions: Handcrafting the Standardized Questionnaire," in *Quantitative Applications in the Social Sciences*. Newbury Park, CA, 1986, pp. 80.

[18] M. P. Couper, "Web surveys: A review of issues and approaches," *Public Opinion Quarterly*, vol. 64, pp. 464-494, 2000.

[19] J. R. Zaller, *The nature and origins of mass opinion*. New York: Cambridge University Press, 1992.