Presenting Diverse Political Opinions: How and How Much

Sean A. Munson **Paul Resnick**
School of Information, University of Michigan
1075 Beal Avenue, Ann Arbor, MI 48109
{samunson, presnick}@umich.edu

ABSTRACT
Is a polarized society inevitable, where people choose to be exposed to only political news and commentary that reinforces their existing viewpoints? We examine the relationship between the numbers of supporting and challenging items in a collection of political opinion items and readers' satisfaction, and then evaluate whether simple presentation techniques such as highlighting agreeable items or showing them first can increase satisfaction when fewer agreeable items are present. We find individual differences: some people are diversity-seeking while others are challenge-averse. For challenge-averse readers, highlighting appears to make satisfaction with sets of mostly agreeable items more extreme, but does not increase satisfaction overall, and sorting agreeable content first appears to decrease satisfaction rather than increasing it. These findings have important implications for builders of websites that aggregate content reflecting different positions.

Author Keywords
Diversity, presentation, design, news, opinion, politics, preferences, selective exposure, news aggregators.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Experimentation, Design

INTRODUCTION
There are many websites that aggregate political news and opinion, such as Digg, Reddit, Slashdot, and Memeorandum. These aggregators present links to recent articles from blogs and commercial media, and many also offer a forum for readers to discuss the linked stories. As the news aggregator marketplace matures, consumers will have many options to choose from and, over time, they will gravitate to aggregators that offer the mix of articles and the type of discussion that they like best.

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

It is easy to gather feedback from users about which articles they like and which articles they agree with politically. But what mix of agreeable and challenging articles would people ideally like to see? From a designer's perspective, there are actually two versions of this question, depending on whether the news aggregator will present each reader a potentially different collection of articles, or whether the same collection will be presented to everyone. If personalized collections will be presented, the question is, "what is the optimal percentage of agreeable items to present?" If the same collection will be presented to everyone, the question is, "is it possible to keep a set of readers with diverse political preferences satisfied or will groups with different political preferences inevitably drift towards using separate aggregator sites?" The latter is a very real possibility, as the launch of several avowedly conservative competitors to Digg (e.g. Lively Links, GOPHub, and the somewhat awkwardly named R-igg) or YouTube (e.g. PopModal.com) might suggest, though none of these have yet gained the popularity of Digg or YouTube.

Providing people with only agreeable news items may have negative social consequences, for two reasons. First, deliberation experiments have shown that interaction with like-minded people leads to polarization: participants tend to end up with more extreme views than they started with [22]. Selective exposure to reinforcing news and opinion articles might also lead to opinion shifts to more extreme positions, and fragmentation of the audience to different sites may lead to discussions of articles that lead to polarization. Increased polarization would make it harder for society to find common ground on important issues. Second, exposure to, and inclusion of, diverse opinions can also lead to more divergent, out of the box thinking, which can improve individual and group problem solving and decision-making [14, 15]. Third, there is a natural tendency for people, particularly those in the minority, to think that their own views are more broadly shared than they actually are [18]. Having a better assessment of their true popularity may lead people to accept the legitimacy of disagreeable outcomes in the political sphere, rather than concocting conspiracy theories to explain how the supposed majority will was thwarted.

Thus, a reasonable public policy goal is for people to be exposed to viewpoints other than their own. Sunstein and others, however, raise alarms that as people have more

choice in news sources, and better tools for filtering out disagreeable news, the opposite will happen [21]. Whether the goal can be achieved in an environment of individual freedom, however, depends on whether designers succeed in selecting single collections that appeal to people with varying political views, or on whether they succeed in creating personalized collections that contain significant challenging information but are liked as much or more than collections without such challenging information.

If designers of news aggregators turn to the research literature to learn how much opinion diversity people seek or tolerate in their political news, they will be confronted with a range of competing theories and evidence. According to different theories, readers may seek out diversity, they may avoid it, or they may seek reinforcement and tolerate challenge only when accompanied by sufficient reinforcing information. Each of these alternatives has different design implications.

Selective exposure theory suggests that people both seek out affirming items and avoid challenging items [5, 13]. We will refer to this as the challenge-aversion hypothesis. Some studies of online political spaces support this homophily theory. Left-leaning and right-leaning blogs rarely link to each other [1], and one study finds that in political blog comments, only 13% of comments expressed disagreement [6]. People who are challenge-averse would prefer news and opinion aggregators that display only agreeable content. If people are challenge-averse, it will be difficult for designers to create single collections that appeal to audiences with diverse opinions. Moreover, in personalized presentations, people will prefer homogeneous collections of all agreeable items. In either case, it will be hard to meet the public policy goal of high exposure to challenging information, unless presentation techniques can be developed that make that exposure more palatable.

Other arguments and studies dispute the challenge aversion theory and offer a contradictory hypothesis of diversity seeking. Sears and Friedman reviewed the literature from the 1950s and 1960s, finding five studies showing a preference for supportive information only, five showing a preference for diversity, and eight inconclusive [19]. Some recent studies support the idea that individuals are using the Internet to seek out a broad range of political opinion and information. In interviews with users of several online political spaces, Stromer-Galley found that those participants sought out diverse opinions and enjoyed the range of opinions they encountered online [20]. A study by the Pew Internet and American Life Project during the 2004 election season found that, overall, Americans were not using the Internet to access only supporting materials [8]. Instead, Internet users were more aware than non-Internet users of a range of political arguments, including those that challenged their own positions and preferences. The preferences of diversity-seeking individuals are consistent with public policy goals; they would be most satisfied with collections that contain a range of views. If people are diversity-seeking, it may be possible to simultaneously satisfy an audience with diverse viewpoints. Moreover, diversity-seeking people would also choose personalized collections with much less than 100% agreeable content.

A third hypothesis is that people are support-seeking. Like challenge-aversion, this hypothesis posits that people seek out affirming items but rejects the idea that they avoid challenging items. In an online experiment, Garrett found that subjects recruited from the readership of left-wing and right-wing sites clicked most on news stories that they expected would reinforce their viewpoints but clicked only slightly less frequently on stories they expected to contain challenging information than on others. Once they looked at those challenging stories, moreover, they tended to spend more time reading them [6]. Garrett proposes a theory, which we call support-seeking, that people seek out supporting items, but are indifferent about challenging items they encounter, so long as they see a sufficient number of supporting items. Support-seeking individuals would prefer news aggregators that show them a sufficient amount of agreeable content, and would be indifferent as to whether items beyond that amount support or challenge their views. Thus, with a large number of items in a collection, it would be possible to please people with a variety of viewpoints, and in an individualized collection, people would not mind the inclusion of additional challenging items.

In this study, we presented readers with lists of political

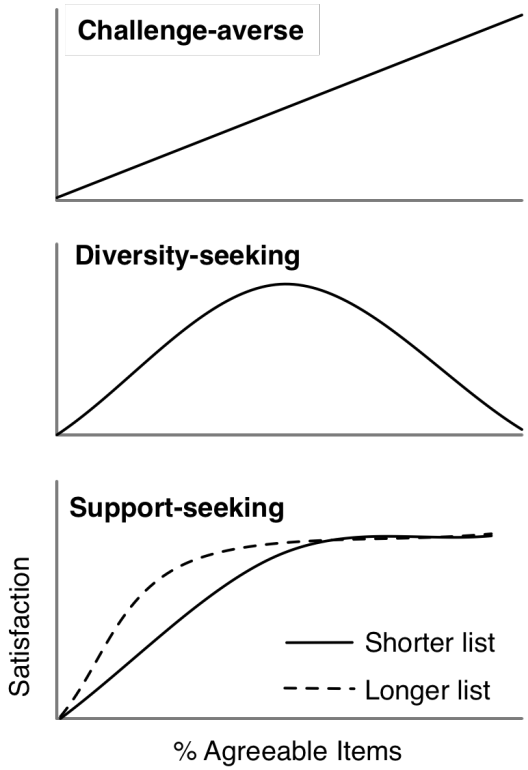

Figure 1. Competing hypotheses about preferences for agreeable and challenging items.

opinion stories, with varying numbers of challenging and agreeable items, and then measured their satisfaction with the list, to distinguish among these different theories about people's preferences for agreeable and challenging information. We would expect to see different relationships between the percentage of the list that is agreeable or disagreeable and reader satisfaction, as summarized in Figure 1. If people are challenge-averse, then we would expect that a higher percentage of agreeable items and lower percentage of disagreeable items would be preferred. On the other hand, if people prefer diversity, then we would expect them to be most satisfied when the list contains both agreeable and disagreeable views. Finally, if people prefer agreeable items but are not particularly averse to challenging items, then the count of agreeable items, not the percent of items that are agreeable, should drive satisfaction. Thus, satisfaction should increase toward the asymptote as the percentage of agreeable items increases, and should approach the asymptote at a lower percentage for longer lists.

If there are people who are challenge-averse or support-seeking, this opens a new question: can we use presentation techniques to make people who prefer support equally satisfied with a lower percent or count of agreeable items? To assess this, we tried highlighting agreeable items, as well as placing them first in the collection.

EXPOSURE TO DIVERSITY

Prior research on exposing individuals to diversity in the news has discussed both the selection and presentation components of the problem.

Park et al developed and evaluated the NewsCube system [17]. NewsCube classifies content into different viewpoints or "aspects" on issues. Subjects who participated in a NewsCube trial read more articles about a controversial issue than subjects who used Google News to learn about the same topic. The articles read by NewsCube subjects appeared to cover a greater breadth of opinions (per reader). Subjects who said they would not normally compare different articles on a topic said that the aspect-presentation made them want to read more articles. This suggests that certain presentations of information can increase the diversity of opinions that participants are motivated to access, even if some of those opinions are disagreeable.

Other work, also examines both the presentation and selection questions. Oh et al built a blog search engine that classified results according to political viewpoints (liberal or conservative) and then used that information in the presentation of search results [16]. Results were either labeled or sorted into two labeled columns according to this bias. Subjects had mixed reactions to these presentations, but generally preferred seeing a column of liberal items and a column of conservative items. Those dissenting preferred to decide for themselves what was liberal or conservative, or felt that the labels added too much polarization to the results – even though they did not change which results were shown. The researchers also found that in the two-column layout, liberal sources accounted for a greater portion of the liberal searchers' clicks (there were too few conservative subjects to observe trends among these subjects).

We previously proposed an algorithm – based on user votes – for selecting diverse items [12]. There, we discussed three metrics for evaluating selection algorithms: inclusion, alienation, and proportional representation. While these metrics provide insight into algorithms' performance at including items for which many users voted, and how well the proportions of items align with proportions of users, they are incomplete as measures of the diversity goals discussed above. That is, while they tell us about whether the collection includes items that represent voters, these metrics do not tell us whether the voters *feel* represented. Many of the desirable and undesirable outcomes depend not on whether diverse results are present but instead on viewers' *reactions* to that diversity. For example, even if ideas are represented in a collection in the same proportion with which they are held, people holding minority opinions may not notice the small number of items representing their views among the many more that challenge their opinions. This may cause them to seek out sources where they can more easily find items that affirm their views, leading to polarization. It is important, then, to understand reactions to different levels of supporting and challenging items, and to identify design choices that may affect viewers' reactions to these collections.

METHODS

All experiments followed the same general design: we provided each subject each day with a list of 8 or 16 links to opinion articles, with known biases, about United States politics. The articles were found on blogs or in the mainstream media. Each item was displayed as an article title together with its first or last paragraph, as shown in Figure 2. The selection and presentation of items varied among subjects.

Subject Recruitment

We recruited subjects from Amazon.com's Mechanical Turk service, a system that allows remote workers to complete small tasks for small payments. Mechanical Turk has been used by other researchers for annotating data, and Kittur et al have published guidance for using Mechanical Turk in research [9]. These guidelines were useful in planning our study.

We used a Qualification test, for which subjects were not paid, to initially screen the Mechanical Turk workers. We asked subjects about their location, age, political knowledge, and political preferences.

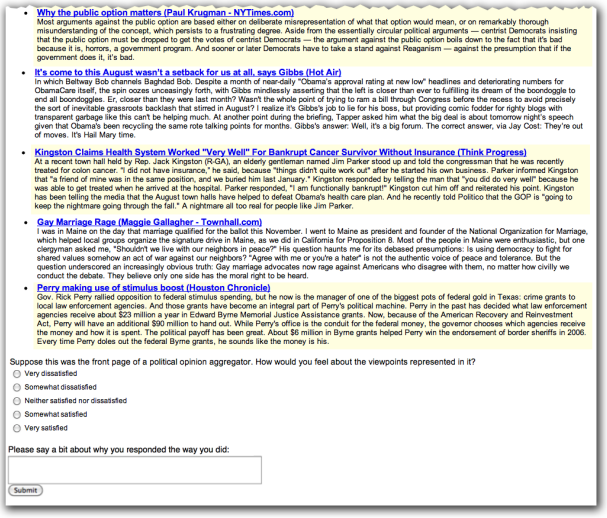

Figure 2. Example articles and question form.

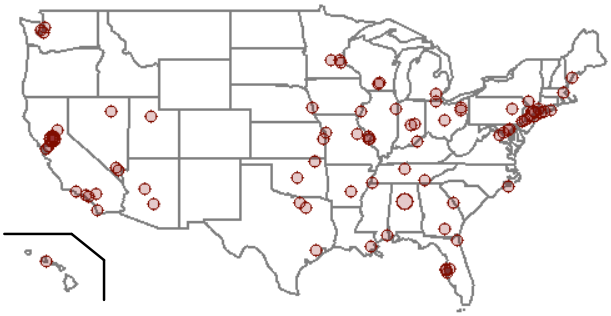

Figure 3. Locations of Mechanical Turk workers who participated in the study.

We only accepted Mechanical Turk workers who self-reported (to Amazon for payment purposes) a United States location and whose previous task approval rate exceeded 90%. Additionally, we screened each Mechanical Turk worker for some basic US political knowledge using multiple-choice questions:

- Who is the current Vice President?
- Which party is George W. Bush a member of?
- For what position is Sonia Sotomayor currently a nominee? (And later, to what position has Sonia Sotomayor been appointed?)

Potential subjects had to correctly answer two of the three questions to participate, though on average, accepted subjects answered 2.98 questions correctly.

We also asked political subjects about their party affiliation (7-point scale from strong Democrat to strong Republican) and about their political preferences (7 point scale from strong liberal to strong conservative). We selected for subjects whose party affiliation matched their liberal or conservative preferences (for example, we screened out subjects who reported being both a strong Republican and a liberal). Furthermore, because we wanted subjects for whom we could predict an article would be agreeable or challenging with some confidence, we also filtered out subjects who were more neutral or independent. To be included, subjects had to have a mean position (as the average of their responses to the two seven-point scale questions) of ≤ 3 or ≥ 5.

Although subjects were not a random sample of the U.S. population, the subjects were diverse in geography, age, and gender. Subjects live in 37 of the 50 U.S. states (Figure 3). Their mean age was 34.3 years (median: 31 years, standard deviation: 11.8 years). 83 were men, 87 women.

Item selection

We selected items based on links from a panel of 500 political blogs. We coded each of the source blogs based on its political ideology (liberal, independent, or conservative). We consulted both Wonkosphere and PresidentialWatch08, which maintain directories of weblogs classified by political affiliation. In addition, one of the authors read entries from each blog and coded the blog manually. When the three classifications disagreed, the majority classification prevailed. If a blog was only classified by one of Wonkosphere and PresidentialWatch, and there was disagreement between that source and the reader, we chose the blogger's self-identification (if present) or the third-party (Wonkosphere or PresidentialWatch) assessment. Our panel of blogs contained 259 liberal blogs (52%), 177 conservative blogs (35%), and 64 independent blogs (13%)

Each day, we selected the 40 most popular liberal and conservative items, based on number of links to the items from liberal and conservative blogs in the previous 36 hours. Items were defined as conservative if the ratio of the probability of any conservative blog in our panel linking to the item compared with the probability of any liberal blog in our panel linking to the item was at least 2:1, and vice-versa for liberal items. The selection system also filtered out tweets, Twitter accounts, Wikipedia articles, and YouTube videos. Before including these articles in our pool of items, each morning researchers manually inspected each candidate item and removed items that did not match the predicted bias (e.g. a liberal item coded as "conservative" because conservative bloggers linked to it to highlight a disagreement with liberals). We also removed items that did not contain or report on opinion, as well as posts that contained only video, images, or audio. On average, this left 23 articles of each bias, per day.

30 turkers were assigned to a manipulation check survey. In this survey, each turker was presented with a list of three links and asked to what extent they agreed with each link on a 5-point scale. This was run early each day, after the researchers reviewed the list of items. Based on raters' responses, we then removed items that were not found to match their predicted bias. An item that was predicted to be liberally biased would be removed if liberal raters did not agree with it, or if conservative raters did agree with it.

Raters agreed with our predicted ratings (as supporting or challenging the raters' opinions) 74% of the time. These disagreements were not distributed evenly across items, and 16% of the items were removed because of disagreement.

We did not require all items to get 100% agreement for inclusion, as we do not expect parties to be 100% united in opinion. Initially, an item was queued for 3 ratings (including at least 1 from a participant from each party). If no raters diverged from the prediction, an item was included. If two or more raters diverged from the prediction, we discarded the item. If one rater diverged from the prediction, the item was queued for an additional rating from someone with the same party affiliation. If the fourth rater disagreed with the prediction, the item was discarded; if this rater agreed, the item was included. In cases when this rating was not obtained (insufficient participation in our Mechanical Turk task), the item was discarded.

On five days, this removal resulted in too few liberal or conservative items, in which case the system filled in with items from the previous day as needed.

Experimental Design

The subjects not assigned to the manipulation check condition viewed a list of items and were asked one of three questions about the representativeness of the collection as a whole. Subjects were randomly assigned to one of six experimental conditions in a 2x3 factorial design:

- *Total number of items:* 8 or 16
- *Presentation:* a list with agreeable and disagreeable items interwoven, a list with agreeable and disagreeable items interwoven and the agreeable items highlighted, or a list with the agreeable items first and highlighted followed by the disagreeable items.

Repeated measures were collected: each subject could complete one survey per day, based on the items selected that day. Subjects remained in the same experimental condition throughout the study, but the number of agreeable items was randomly chosen for each subject each day.

The instructions read:

The following list contains some of the most-linked to political opinion stories from the last few days. Please look at the list as you might if you were to visit a website like Digg or Reddit (you may click on and read as many or few as you like). Then answer the questions at the bottom of the page. Thank you!

Additionally, subjects in the agreeable first or highlight conditions were told that items they were predicted to agree with would appear highlighted in the list.

Subjects were also assigned to one of two questions. The first question (44 subjects) was our primary outcome measure and asked about the subject's satisfaction with the range of views:

"Suppose this was the front page of a political opinion aggregator. How would you feel about the viewpoints represented in it?" (5 point Likert scale, Very dissatisfied to very satisfied).

The second question (39 subjects) asked about the bias of the collection:

"What, if any, is the political bias of this collection?" (5 point Likert scale, Very Liberal to Very Conservative)

We used this question to help us understand our findings. If a particular presentation feature affected satisfaction, is it because it changed how the subjects perceived the collection's bias, or is it because the subjects simply did not like the feature?

In addition to these questions, each time a subject viewed a list, they were randomly asked either to provide a free-text explanation for why they gave the rating they did or to repeat a question from the pre-test (party affiliation, liberal to conservative, age, or gender). The free response question helped us to understand why subjects gave the ratings they did and if they were interpreting the questions as intended. Repeating questions from the qualification test follows a recommendation from Kittur et al to ask verifiable questions [9]. 5 subjects (4 from the satisfaction question, 1 from the bias question) changed their answer substantially (e.g. aging more than one year or in reverse, changing gender, or shifting on either of the political spectrum questions by 2 points or more). Though there are many possible explanations for these shifts – such as shared accounts within a household, careless clicking, easily shifting political opinions, deliberate deception, or lack of effort – all of these explanations are not desirable for study subjects, and so these subjects and their responses were excluded from our analysis, leaving us with 108 subjects (40 responding to the satisfaction question, 38 responding to the bias question, and 30 rating articles for the manipulation check).

After a five-day warm-up period with variable pay (to identify an appropriate price), subjects were paid $0.75 for rating a collection of items. This pay may seem high compared to some expectations for Mechanical Turk labor. We believe that we had to pay a higher price because our task both required successful completion of a qualification and was only available once per day. Interestingly, many more people (171) completed the qualification test and were approved than actually returned to complete a task.

We collected data daily from 22 July – 14 August, and then on alternating days from 26 August to 10 September, 2009. This time period and major topics in articles displayed included national debates about healthcare reform and the Waxman-Markey cap and trade bill, the death of Massachusetts Senator Edward Kennedy, discussion about the success of the Troubled Asset Relief Program and the American Recovery and Reinvestment Act (including the "Cash for Clunkers" program), the release of two American

journalists from North Korea, continued discussion about political unrest in Honduras, the end of Alaska Governor and former Vice Presidential candidate Sarah Palin's term, and speculation about 2010 Congressional and gubernatorial elections.

A previous study found that Mechanical Turkers' efforts were not tied to the amount of payment [11]. On average, our subjects rated a list in 6.3 minutes (5.8 minutes for 8-item lists; 7.3 minutes for 16-item lists). No subjects completed the task in less than a minute. For comparison, Alexa (accessed 16 September) reports that Digg visitors spend an average of 4.2 minutes per day on the site, Reddit visitors spend an average of 6.3 minutes, and Memeorandum visitors spend an average of 2.0 minutes.

RESULTS

Diversity preferences

In our first look at the data, we found that when the list contained a low percentage of agreeable items, almost all subjects were very dissatisfied. When a high percentage of items were agreeable, however, there was greater variance in responses: some subjects were very satisfied, some subjects were very dissatisfied. This suggested that there may be individual differences, with some people diversity-seeking and some challenge-averse.

To confirm this, we analyzed the open-ended responses for the question about why subjects in the satisfaction question group gave the ratings they did. Some subjects wrote that they specifically did not want a list of solely supportive items and that they want opinion aggregators to represent a fuller spectrum of items, even if that includes challenge. We coded all free-text responses for similar remarks, and then coded participants as diversity-seeking if they had made at least one such comment. Our standards were strict: subjects had to write that they either

1. wanted a fuller spectrum of views even though their views were represented in the majority of items in the list, e.g.

   "It all seems liberal. I'm liberal, but I think it's good to get dissenting opinions instead of having all the articles slanted the same way. I'd really like seeing pro and con articles on some of the topics."

   "The articles in this list showed some of both sides on some issues, but on other issues like health care was rather one sided. If that and a few other articles had been given two sides I would be completely satisfied. I like to read both sides even though I am mostly conservative."

   or

2. were pleased with the balance of items and would not want more supporting items, e.g.

   "There is an even distribution of right and left wing articles. I think it is best to cover both sides of the issue."

   "I like that there are views from both Democrats and Republicans and seems to be a great mix of both sides of the fence."

To avoid potential biasing of the coders, coders were not informed of the actual number of agreeable items presented to a subject or of the subject's satisfaction score for the items when coding the subject's explanatory comments. We did not use the actual properties of the list associated with a comment when we coded, out of concern that we would code people as diversity seeking when the subject's remarks were ambiguous in order to explain behavior. Our inter-rater reliability, calculated as Cohen's kappa [4], was 0.89. All raters coded all subjects. Landis and Koch characterize agreement above 0.8 as "almost perfect agreement" [5]. We decided the disagreements through discussion.

10 out of the 40 subjects in the satisfaction condition were coded as diversity-seeking (25%). This is likely an undercount given our coding criteria and that some participants never saw a collection that would prompt different reactions from diversity seeking participants. Once

	β	Std Err	p-value
Intercept	1.30	0.23	<0.001
% Agreement	2.28	0.76	<0.010
(% Agreement)^2	0.80	0.66	ns
Diversity seeking	-0.25	0.63	ns
% Agreement * Diversity seeking	6.49	3.16	<0.050
(% Agreement)^2 * Diversity seeking	-8.32	3.11	<0.050

Table 1. Linear regression results for satisfaction (1-5). n=145 from 40 subjects, $F(5,39) = 29.63$ ($p < 0.001$); adjusted R^2 0.4776.

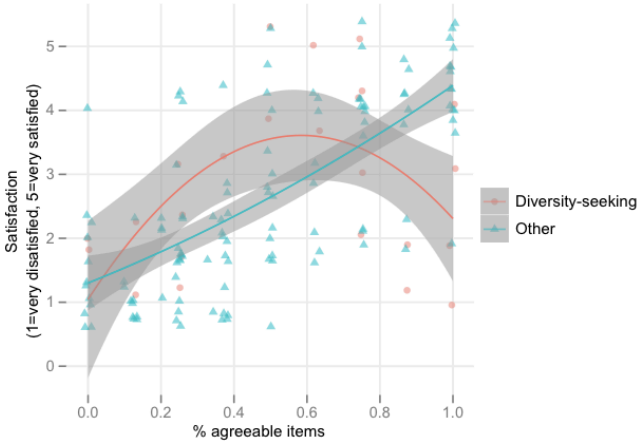

Figure 4. Comparison of satisfaction at different percentages of agreeable items for diversity-seeking and other (either challenge-averse or support seeking) individuals. Fit lines according to regression model in Table 1. The grey band includes ± one standard error of the prediction.

we separated out the diversity-seeking individuals, our results were much clearer (Figure 4). The linear regression with percent agreeable items, whether a subject is diversity seeking, and the interaction terms between these (to the quadratic polynomial) shows significant interaction effects between percent agreement and diversity seeking for predicting satisfaction (Table 1). Because subjects were able to rate multiple collections (one per day), in all regression results we cluster responses by subject, which reduces degrees of freedom and inflates standard error estimates, to correct for correlation of repeated measures.[1]

Support-seeking vs. Challenge-averse

We then examined whether the remaining 30 individuals were support-seeking or challenge-averse. Challenge-averse subjects would be equally satisfied at the same percentage of agreeable items, regardless of the length of the list. Support-seeking subjects, in contrast, would be equally satisfied at the same number of agreeable items, regardless of the length of the list. If we were to find evidence that people are support-seeking, it could be possible to address the public policy challenge of exposing people to more perspectives, in the proportions with which they are held by the population, simply by presenting a longer list of results.

In our analysis, we did not find evidence of support-seeking individuals. Table 2 presents the linear regression model for:

$$\text{satisfaction} = \beta_1 + \beta_2(\%\text{ agreeable items}) + \beta_3(\text{listlength}_{16}) + \beta_4(\text{listlength}_{16} * \%\text{ agreeable items})$$

In this model, listlength_{16} is a dummy variable equal to 1 for 16-item lists (longer lists) and 0 for 8-item lists (shorter lists).

We exclude data from the 10 diversity-seeking individuals. If there are any effects of list length, they are in the opposite direction of what would be expected for support-seeking behavior. In a plot of the percent agreeable items and satisfaction (Figure 5, top), the slope of the fit lines for the two list lengths follow each other quite closely, suggesting that count does not matter. When we plot the number of agreeable items (Figure 5, bottom), we can see a clear divergence. Furthermore, 2 agreeable items out of a total of 8 is superior to 2 agreeable items out of a total of 16 ($t(7.373) = 3.3471$, $p<0.05$). Clearly, the presence of challenging items, not just the count of agreeable items, drives satisfaction. We conclude that the remaining subjects as a group are challenge-averse, though a few individuals may be support-seeking.

Presentation Techniques for Challenge-averse Readers

Having found that at least some people exhibit challenge-averse behavior, we investigated whether presentation techniques could increase their satisfaction with collections that contained some challenging items.

	β	Std Err	p-value
Intercept	1.61	0.33	<0.001
% Agreement	2.69	0.40	<0.001
List length = 16	-0.58	0.40	ns
% Agreement * (List length = 16)	0.53	0.53	ns

Table 2. Linear regression results for satisfaction (1-5) with data from diversity-seeking subjects withheld. $n = 112$ from 30 subjects, $F(3,29) = 50.43$ ($p < 0.001$); adjusted R^2 0.5079.

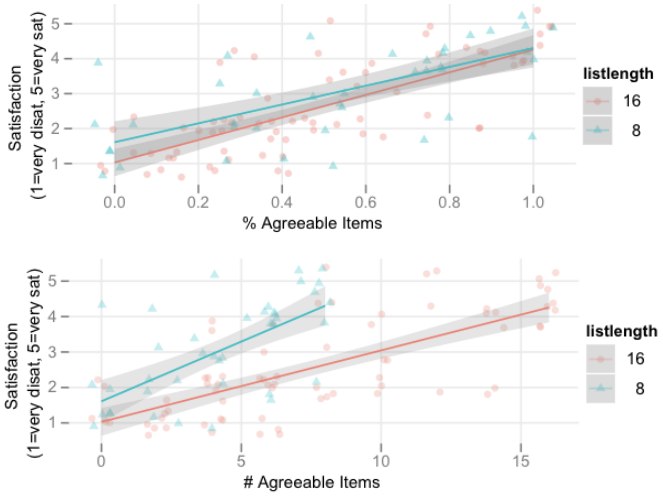

Figure 5. *Top*: Percent agreement and satisfaction for 8 and 16 item lists. Fit lines from model in Table 2. *Bottom*: Number of agreeable items and satisfaction for 8 and 16 item lists. Fit lines: satisfaction = β_1 + β_2(# agreeable items) + β_3(list length)+ β_4(list length * # agreeable items). Responses from individuals coded as diversity-seeking excluded from both plots.

Table 3 presents a linear regression model for the effects of presentation style and their interaction effects with percent agreement. We will discuss the effects of each presentation style separately.

Highlighting Only

We had expected that, for challenge-averse individuals, highlighting agreeable items would increase their satisfaction at all percentages of agreeable items, by helping them identify these items even when they were rare in the collection. Figure 6 presents results from a reduced model excluding participants in the agreeable first condition.

Contrary to our expectations, there is no main effect of highlighting. Instead, there is a significant interaction terms between highlighting and percentage of agreeable items. Highlighting agreeable items makes a reader's reaction – whether it is satisfaction with a high percentage of agreeable items or dissatisfaction with a low percentage – more extreme.

At the lower range of agreeable items, where we had expected highlighting have the greatest improvement in

[1] All regressions were performed using STATA 10's `regress` command with the robust cluster by subject option.

	β	Std Err	p-value
Intercept	1.59	0.29	<0.001
% Agreement	2.60	0.36	<0.001
Highlighting only	-0.60	0.41	ns
% Agreement * Highlighting only	1.29	0.60	<0.05
Agreeable first	-0.97	0.31	<0.010
% Agreement * Agreeable first	0.64	0.44	ns

Table 3. Regression model for a reader's satisfaction (1-5_ as predicted by percent agreement and presentation style. (baseline presentation interweaves agreeable items and does not include highlighting). $n = 121$ from 30 subjects, $F(5,29) = 67.42$, $p < 0.001$, adjusted $R^2 = 0.564$.

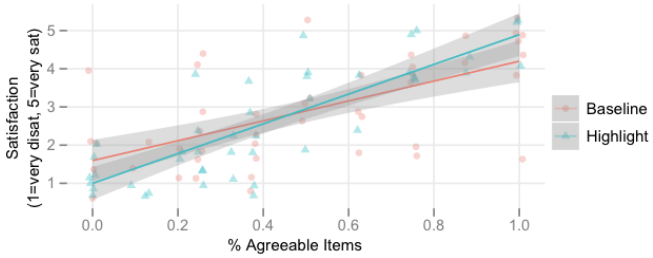

Figure 6. Reduced model for highlight -only:
satisfaction = $\beta_1 + \beta_2$(% agreeable items) + β_3(highlight) + β_4(highlight*% agreeable items)

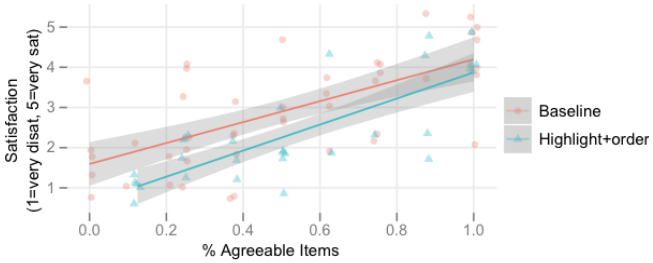

Figure 7. Model comparing baseline and highlight+ordering:
satisfaction = $\beta_1 + \beta_2$(% agreeable items) + β_3(agreeable first) + β_4(% agreeable items * agreeable first)

satisfaction, subjects' satisfaction is actually decreased. In the mid-range, particularly important for public policy goals of showing viewpoints in proportion to how they are held, highlighting has essentially no effect. With high percentages of agreeable items, it may be possible to highlight agreeable items and include a few challenging items while achieving the same satisfaction as a list of only agreeable items.

Among subjects to whom we posed the question about the collection's bias, we saw a similar interaction effect between highlighting and the percent of agreeable items, suggesting that the effects of highlighting on a subject's satisfaction with the collection is moderated by how highlighting affects their perception of bias in the collection. In other words, it seems that highlighting helps subjects judge the percentage of agreeable items, and the perceived percentage drives satisfaction. Consistent with that interpretation, we note that challenge-averse readers in the highlight condition spent an average of 5.1 minutes per collection, compared to 7.5 minutes for challenge-averse readers who read lists without highlighting.

Highlighting+Ordering: Agreeable items first

We also anticipated that placing agreeable items first would increase challenge-averse readers' satisfaction. Instead, readers who viewed agreeable items as highlighted and at the beginning of the collection reported lower satisfaction (Table 3). Figure 7 displays a model comparing highlight+ordering and the baseline presentation.

The subjects who we asked to rate the collection's bias, however, reported that the collection was more biased in their favor when the agreeable items were highlighted and shown first than when the agreeable items were not highlighted and were interwoven with disagreeable items. This appears to be contradictory, or at least suggests that something other than perceived bias is driving satisfaction in this case. Challenge-averse readers in the highlight+ordering condition spent an average of 5.3 minutes per collection, compared to 7.5 minutes for challenge-averse readers who read lists without ordering and highlighting.

DISCUSSION

Our study finds good and bad news for those with a policy goal of encouraging exposure to a diversity of opinion. The good news is that some people actually prefer collections of items with diverse opinions. They appear not to be the majority, and so it may be important to consciously design specifically for this audience, as they may not naturally be served if designers build applications primarily for the majority, challenge-averse individuals.

The bad news is that for challenge-averse individuals, designers cannot substitute ordering or highlighting of agreeable items for including more agreeable content. With highlighting, it might be possible to include one or two challenging items in a list of otherwise agreeable items and achieve the same satisfaction from challenge-averse people as with an unhighlighted list that happens to contain completely agreeable items. From the perspective of website operators trying to attract and retain users, this is unlikely to be a desirable tradeoff. It is unlikely to be sufficient challenge to satisfy diversity-seeking individuals, and would leave them vulnerable to losing challenge-averse individuals to competitors who offer 100% agreeable items all the time (and hence need no highlighting).

We also cannot rule out that the observed effect of placing agreeable items first is a result of flawed experimental design. By asking about subjects' satisfaction at the end, and placing the agreeable items at the beginning, we may have prompted a recency effect [2] – that is, their answer was more influenced by the disagreeable items nearer to the question. In an actual political opinion aggregator, truly

challenge-averse readers may never scroll that far, while the Mechanical Turk readers may have felt an obligation to read every item because they were being paid rather than reading for their own enjoyment, or they may have skipped looking at the first few items, immediately scrolling down to the questions and looking only at the items closest to the questions.

The responses from the subjects who we asked about the collection's bias, however, appear to contradict the explanation that items nearer the bottom of the list, and thus the question, weighed more heavily in the subjects' consideration of bias and led to the observed decrease in satisfaction with presenting agreeable items first. It is possible that the subjects simply liked the agreeable first presentation less. It is also possible that subjects read the collection differently when we asked them to characterize its bias than when we asked them about their satisfaction with the opinions presented, and that this caused them to experience the collection's bias differently between the two subject groups.

Alternative experimental designs – such as placing the question at the top of the list or having it scroll alongside – may lead to an improved understanding of the effect of ordering. It is also possible that subjects in a lab experiment or Mechanical Turk task will always feel obligated to read an entire list even when they would not do so on an actual website. Field experiments may be necessary to study item ordering.

We only evaluated a small range of the presentation techniques for lists of items. Alternative ideas that could be evaluated include reducing the space that challenging items occupy in a list by reducing the font size or collapsing challenging items' abstracts. Perhaps such techniques would make palatable a smaller percentage of agreeable items. They would increase the risk, however, that people would not actually be exposed to the challenging items, thus thwarting the public policy goal of increasing exposure to diversity.

More sophisticated presentation techniques, such as aspect browsing in NewsCube [17], may have more potential for showing diverse items to challenge-averse individuals. Agreeable items might be shown on the front page, with challenging items on the same topic linked from the agreeable item page. A similar idea might be based on the presentation used by the political news aggregator Memorandum. Memeorandum groups items by topics. The front page includes abstracts for top items with links to other items on the same topic below the abstract. To appeal to diversity-seeking individuals, the display might be personalized to show a top-level item and abstract for any topic from a supportive source, with more challenging items appearing in the links.

Another limitation of this study is that we have reduced the political spectrum to two broad points of view, and that our subjects only include people whose views fit at the ends of this axis. We do not know if our results generalize to people who are less partisan or more independent. This limitation is another reason that our estimate of the percentage of diversity-seeking individuals must be taken with skepticism.

Future work should further explore the distribution of individuals' preferences for diversity. Though we find that at least some people are diversity-seeking and at least some people are challenge-averse in their preferences for political news and opinion, we do not know the distribution of preferences in the population as a whole, or if an individual's preferences are common across topic areas, or if someone who prefers to avoid challenging political opinions may seek out challenging opinions about which baseball team will win the pennant this year. Given data about which articles people click on, it might be worthwhile to formulate alternative stopping rules, analogous to those hypothesized in the arena of information acquisition for decision-making or design, and to estimate their prevalence or the conditions under which people make use of different stopping rules [3].

Future research also should go beyond short-term measurements and emphasize longitudinal studies. Preferences may vary quite a bit with long-term use of a news aggregator. For example, diversity seekers might prefer diversity one day but tire of it in the long run. Similarly, people who currently are getting diversity might be happy with all agreeable items in the short term, but may not want only supportive items in their day-to-day news source. Other factors, such as whether one's political party is in power, may also affect an individual's diversity preferences over time.

CONCLUSION

In this study, we find a possible reconciliation of the conflicting theories of diversity-seeking and challenge avoidance: they correctly describe the preferences of different groups of people. Contrary to the implicit assumptions of previous research on selective exposure, neither diversity-seeking nor challenge-avoidance is a fundamental trait of human behavior that describes all people.

The presentation techniques of sorting and highlighting were not very helpful at making challenge more appealing to the challenge-averse people, except that highlighting may make a very small percentage of challenging content palatable. Future work should study additional presentation techniques, including more sophisticated displays of challenging and supporting content. Also, rather than trying to increase the percentage of challenging information in the collections shown to challenge-averse readers, it may be more effective to serve well the needs of those who are diversity seeking and provide them with the means to spread insights they gain from challenging content to the people who avoid such exposure in their everyday news reading.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under award IIS-0916099. We are grateful for feedback on the experimental design from participants in the Behavioral and Experimental Economics and Incentive Centered Design lab groups at the School of Information. We thank members of the School of Information's HCI Writers' Group and several anonymous reviewers for helpful critique of earlier versions of this paper.

REFERENCES

1. Adamic, L. and Glance, N. (2005). The Political Blogosphere and the 2004 US Election: Divided They Blog, *Proc. 3rd international workshop on Link Discovery*, pp. 36-43.
2. Broadbent, D.E., and Broadbent, M.H.P. (1981). "Recency Effects in Visual Memory," *Quarterly Journal of Experimental Psychology* 33(A): 1-15.
3. Browne, G.J. and Pitts, M.G. (2004). "Stopping rule use during information search in design problems," *Organizational Behavior and Human Decision Processes*, 95(2): 208-224.
4. Cohen, J. (1960). "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement* 20(1): 7–46.
5. Frey, D. (1986). "Recent Research on Selective Exposure to Information," *Advances in Experimental Social Psychology* 19: 41-80.
6. Garrett, R. K. (2009). "Echo chambers online?: Politically motivated selective exposure among Internet news users," *Journal of Computer Mediated Communication*, 14(2), 265-285.
7. Gilbert, E., Bergstrom, T., and Karahalios, K. (2009). "Blogs Are Echo Chambers: Blogs Are Echo Chambers," *Proc. of HICSS 2009*.
8. Horrigan, J., Garrett, K., and Resnick, P. (2004). "The Internet and Democratic Debate," Pew Internet and American Life Project, October 27, 2004.
9. Kittur, A., Chi, E.H., and Suh, B. (2009). "Crowdsourcing User Studies With Mechanical Turk," *Proc. CHI* 2009: 453-456.
10. Landis, J.R. and Koch, G.G. (1977). "The measurement of observer agreement for categorical data," *Biometrics* 33: 59-174.
11. Mason, W. and Watts, D.J. (2009). "Financial incentives and the 'performance of crowds,'" *SIGKDD Workshop on Human Computation*: 77-85.
12. Munson, S.A., Zhou, D.X., and Resnick, P. (2009). "Sidelines: An Algorithm for Increasing Diversity in News and Opinion Aggregators," *Proc. ICWSM* 2009.
13. Mutz, D.C. and Martin, P.S. (2001). "Facilitating Communication Across Lines of Political Difference: The Role of Mass Media," *American Political Science Review* 95(1): 97-114.
14. Nemeth, C.J. (1986). "Differential contributions to majority and minority influence," *Psychological Review* 93(1): 23-32.
15. Nemeth, C.J. and Rogers, J. (1996). "Dissent and the search for information," *British Journal of Social Psychology* 35: 67-76.
16. Oh, A., Lee, H., and Kim, Y. (2009). "User Evaluation of a System for Classifying and Displaying Political Viewpoints of Weblogs," *Proc. ICWSM* 2009.
17. Park, S., Kang, S., Chung, S., and Song, J. (2009). "NewsCube: delivering multiple aspects of news to mitigate media bias," *Proc. CHI 2009*.
18. Sanders, G.S. and Mullen, B. (1982). "Accuracy in perceptions of consensus: Differential tendencies of people with majority and minority positions," *European Journal of Social Psychology* 13(1): 57-70.
19. Sears, D.O. and Friedman, J.L. (1967). "Selective Exposure to Information: A Critical Review," *Public Opinion Quarterly* 31(2): 194-213.
20. Stromer-Galley, J. (2003). "Diversity of Political Opinion on the Internet: Users' Perspectives," *Journal of Computer-Mediated Communication* 8(3).
21. Sunstein, C. (2001). *Republic.com*. Princeton University Press, Princeton, NJ.
22. Sunstein, C. (2002). "The Law of Group Polarization," *The Journal of Political Philosophy* 10(2): 175-195.