Improving the Presentation and Interpretation of Online Ratings Data with Model-Based Figures

[Information Science]

Daniel E. HO and Kevin M. QUINN

Online ratings data are pervasive, but are typically presented in ways that make it difficult for consumers to accurately infer product quality. We propose an easily understood presentation method that has the virtue of incorporating a parametric model for the underlying ratings data. We illustrate the method with new data on the content quality of news outlets, and demonstrate its reliability and robustness with an experiment of online users and a simulation study. Our simple approach is easy to implement and widely applicable to any presentation of ratings data.

KEY WORDS: Ordinal item response; Posterior simulation; Statistical graphics.

1. INTRODUCTION

The Internet has deluged consumers with information. One prevalent category of such information consists of ratings from Internet users evaluating any number of goods from music, movies, restaurants, companies, and consumer products, to legal cases, and even professors. Table 1 presents some examples from prominent Web sites, such as Amazon and Netflix, to more niche-based Web sites, such as Transport Reviews (posting ratings of auto transporters) and Rate My Professors (posting ratings of college professors). Needless to say, such new sources of information hold great promise for the sophisticated consumption of goods and services. Yet with this explosion of information, fundamental statistical principles of data analysis and presentation are often ignored, thereby making the proper substantive interpretation of information difficult, if not impossible.

Current graphical displays of online ratings data are lacking in three respects. First, current practice ignores *systematic differences across raters*. Some raters may be more inclined to use high ratings while others may use all the rating categories but do not discriminate very well between truly high and low quality products. Weighting all raters equally regardless of what is known of their overall rating behavior (as is generally current practice) has the potential to bias results and makes it potentially easy for users to manipulate overall ratings. Second, current practice fails to incorporate *statistical uncertainty* in the ratings. Users typically observe only a discrete number of stars, as illustrated in the right column of Table 1. Some, but not all, Web sites attempt to cure this problem by presenting the total number of ratings submitted. Although this provides some measure of uncertainty to users, sampling variability alone is not the only reason to be uncertain of true product quality. Finally, most Web sites unnecessarily *discretize* mean ratings by rounding them to the nearest whole number, thereby discarding potentially valuable information. Some Web sites, such as Netflix, provide partial stars, but even this practice remains somewhat uncommon.

We address this problem by proposing model-based graphical displays that: (1) adjust for rater-specific factors, (2) are easily interpretable, and (3) are more accurate than extant graphical displays.

2. AN EXAMPLE DATASET: MONDO TIMES NEWS OUTLET RANKINGS

As a running example, we use data from Mondo Times (*http://www.mondotimes.com/*), an online company that disseminates information about media outlets. Mondo has more than 60,000 users and contains information on 16,920 newspapers, magazines, radio stations, and television stations in 211 countries. Raters submit five-point ratings of the content quality of news

Daniel E. Ho is Assistant Professor of Law and Robert E. Paradise Faculty Fellow for Excellence in Teaching and Research, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA 94305 (E-mail: *dho@law.stanford.edu*). Kevin M. Quinn is Associate Professor, Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (E-mail: *kevin_quinn@harvard.edu*). We thank Eric Kallgren, Mondo Code LLC, and Mondo Times for providing us with data. We also thank the Editor, an Associate Editor, three anonymous referees, Adam Glynn, Eric Kallgren, Gary King, and Tim Shapiro for helpful comments on an earlier draft; Migara De Silva for helpful background information; Troy DeArmitt, Jim Greiner, Justin Hansen, Jon Hennessy, Jen Keighley, Mike Kellermann, Pete Mandel, Andrew Martin, Ellie Powell, Tim Shapiro, and Garth Sheldon-Coulson for comments on the survey design; and Shaun Wolf Wortis for his work on survey implementation. We graciously acknowledge the financial support of the National Science Foundation (grants SES 03-50613 and SES 07-51834). Quinn thanks the Center for Advanced Study in the Behavioral Sciences for its hospitality and support. A version of this article with more details is available at the authors' Web sites (*http://dho.stanford.edu/* and *http://www.people.fas.harvard.edu/ kquinn/*).

©2008 American Statistical Association DOI: 10.1198/000313008X366145

outlets from awful, poor, average, very good, to great. Only Mondo Times members are eligible to submit ratings; membership requires a valid e-mail address, and multiple submissions from the same IP address for one outlet are prohibited.

The dataset used in this article features 1,515 products (news outlets) and 946 raters. Each product was rated between 1 and 89 times and each rater rated between 1 and 130 products. The average number of ratings for a product is 3.0 and the average number rated by a rater is 4.8. The number of ratings in each category summed over all products and raters is: awful, 1003; poor, 606; average, 834; very good, 892; great, 1176. These data are available from the authors and the `Ratings` package (available at *http://cran.r-project.org/*). We use these data not in an effort to conduct a comprehensive analysis, but to illustrate how our proposed methods can be applied to a real dataset.

3. A PRACTICAL SOLUTION: MODEL-BASED FIGURES

The starting point for the statistical graphics discussed in this article is an item response theory (IRT) model for the underlying ratings. IRT models permit researchers to account for systematic inter-rater differences, calculate direct quantities of interest, and estimate associated statistical uncertainty (Bock and Lieberman 1970; Lord 1980; Albert 1992). Interest in and development of these types of measurement models has increased considerably over the past 10 years (Bradlow and Zaslavsky 1999; Patz et al. 2002; Boeck and Wilson 2004; Segall 2004). But while the science of analyzing such data has advanced, the presentation of results has not been well-adapted to the context of Web site ratings—even in spite of the recognition that data visualization remains central to the dissemination and understanding of scientific results (Tukey 1977; Tufte 1983; Cleveland and McGill 1985, 1987; Cleveland 1993; American Association for the Advancement of Science 2004). Although some graphical displays of IRT parameters require users to possess a fair amount of statistical knowledge, our proposed graphs make use of the posterior predictive distribution and are thus easily interpreted as the (posterior) probability of a randomly chosen user providing a particular rating for a product. Further, these graphs require only minor changes to the way that current ratings data are displayed on most Web sites.

3.1 A Statistical Model for Ratings Data

The model upon which our graphs are built is an ordinal item response theory (IRT) model with fixed cutpoints. It is a special case of the models discussed by Johnson and Albert (1999) and Quinn (2004). Because the proposed graphical displays depend only on the posterior predictive distribution of the observed ratings, any reasonable statistical model for these ratings could be substituted. The setup of our model is the following.

Let R denote the set of individuals who have rated at least one product and let P denote the set of products that have at least one rating. To ensure comparability, it is also necessary to ensure that the products are connected in the sense that any two products can be bridged by a sequence of raters who have rated at least some of the same products. In what follows, we use r and p to index raters and products, respectively.

The observed data are the collection of ordinal ratings of the items in P by the raters in R. We let $\mathbf{Y}$, with typical element y_{pr}, denote the collection of all possible $|P| \times |R|$ ratings and y_{pr}^{obs} denote the observed rating of product p by rater r. In many cases, r will not rate p and y_{pr} is not observed. It is assumed that y_{pr} is ordinal and can take values in $\{1, 2, \ldots, C\}$ for all $p \in P$ and $r \in R$. In what follows, we will think of the numbers $1, 2, \ldots, C$ as increasing in quality.

We assume that elements of $\mathbf{Y}^{\text{obs}}$ are generated according to:

$$y_{pr}^{\text{obs}} = \begin{cases} c & \iff y_{pr}^{*} \in (\gamma_{c-1}, \gamma_c] \quad \text{and} \quad z_{pr} = 0 \\ \text{missing} & \iff z_{pr} = 1, \end{cases}$$

where y_{pr}^{*} is a latent variable, z_{pr} is a missingness indicator, and $\gamma_0 < \gamma_1 < \cdots < \gamma_C$ are a series of cutpoints. Here we constrain $\gamma_0 = -\infty$, $\gamma_1 = 0$, and $\gamma_C = \infty$ and estimate the remaining cutpoints.

We parameterize y_{pr}^{*} as

$$y_{pr}^{*} = \alpha_r + \beta_r \theta_p + \epsilon_{pr}, \quad \epsilon_{pr} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \quad p \in P, r \in R.$$

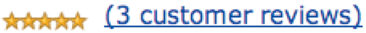
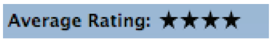
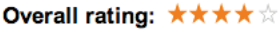
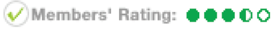
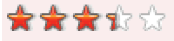
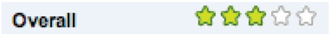
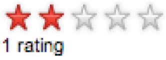
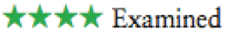
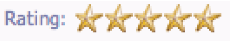
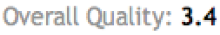
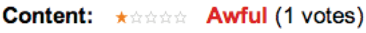
α_r is a scalar parameter that can take values anywhere on the real line and captures the location that rater r uses as the center of her internal scale. If rater r tends to give more negative ratings to all products than other raters, then α_r will be less than the α's of the other raters (i.e., the rater is "critical", similar to users in the left panel of Figure 2). On the other hand, if rater r tends to give more positive ratings to all products than other raters, α_r will be larger than the α's of the other raters. β_r is a scalar parameter that captures how well rater r discriminates between low and high quality products. To identify the model we assume that $\beta_r \in \mathbb{R}_+$ for all $r \in R$. A value of β_r near 0 implies that rater r is unable to distinguish between low and high quality products—her observed ratings are essentially independent of true quality. A large value of β_r implies that r is extremely sensitive to small differences in quality (i.e., the rater is "discriminating," similar to users in Figure 3). θ_p is a scalar parameter that captures the latent quality of product p. With the constraint that $\beta_r \in \mathbb{R}_+$ for all $r \in R$, the interpretation is that quality is increasing in θ_p for all $p \in P$. We assume (a) that data are missing at random and (b) distinctness of model and missingness parameters in the sense of Little and Rubin (1987) and Bradlow and Thomas (1998). In Section 4.2, we explore violations of these assumptions.

We adopt a Bayesian approach to fit this model. The information above is sufficient to write down the sampling density $p(\mathbf{Y}^{\text{obs}}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ for this model as

$$p(\mathbf{Y}^{\text{obs}}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{\{p,r \,:\, z_{pr}=0\}} \left\{ \Phi\left(\gamma_{y_{pr}^{\text{obs}}} - \alpha_r - \beta_r\theta_p\right) - \Phi\left(\gamma_{y_{pr}^{\text{obs}}-1} - \alpha_r - \beta_r\theta_p\right) \right\},$$

where the notation $\gamma_{y_{pr}^{\text{obs}}}$ refers to γ_c if and only if $y_{pr}^{\text{obs}} = c$ and $\Phi(\cdot)$ is the standard normal distribution function. Specification of a joint prior density $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ allows us to write the posterior density $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{Y}^{\text{obs}})$ using the usual formula. For the analysis in the article we assume that each α_r

Table 1. Sample of online ratings data.

Website	Example
Amazon	 ★★★★★ (3 customer reviews)
iTunes	 Average Rating: ★★★★
Epinions	 Overall rating: ★★★★☆
PC Magazine	 ✓ Members' Rating: ●●●◐○
Netflix	 ★★★½☆
Yahoo Shopping	 Overall ★★★☆☆
YouTube	 ★★☆☆☆ 1 rating
Westlaw	 ★★★★ Examined
Transport Reviews	 Rating: ★★★★★
Rate My Professors	 Overall Quality: 3.4
Mondo Times	 Content: ★☆☆☆☆ Awful (1 votes)

follows a normal prior distribution with mean 1 and variance 1, each β_r follows a normal prior distribution with mean −5 and variance 20 truncated to the positive half of the real line, the elements of γ have improper uniform prior distributions, and that each θ_p follows a standard normal prior distribution with all parameters mutually independent. We fit the model using Markov chain Monte Carlo (MCMC). Easy-to-use, free software to implement this model and our proposed figures is available at *http://cran.r-project.org/* as the `Ratings` package.

3.2 Easily Interpretable Figures Based on the Posterior Predictive Distribution

The plots proposed in this article make use of the probability that a randomly selected rater will provide a given rating for a particular product. Because most raters rate very few products, a model-based method to calculate these probabilities is desirable. Our model-based calculations rely on posterior predictive probabilities (Meng 1994; Gelman et al. 1996, 2003; Gelman 2004).

For the model above, the posterior predictive density for y_{pr} is:

$$p(y_{pr}^{\text{rep}}|\mathbf{Y}^{\text{obs}}) = \int\int\int\int p(y_{pr}^{\text{rep}}|\alpha_r,\beta_r,\theta_p,\boldsymbol{\gamma}) \times p(\alpha_r,\beta_r,\theta_p,\boldsymbol{\gamma}\,|\mathbf{Y}^{\text{obs}})d\alpha_r d\beta_r d\theta_p d\boldsymbol{\gamma}\,, \quad (1)$$

where "rep" indicates that y_{pr}^{rep} is a replicated datapoint.

Given M Monte Carlo samples $\left\{\alpha_r^{(m)}, \beta_r^{(m)}, \theta_p^{(m)}, \boldsymbol{\gamma}^{(m)}\right\}_{m=1}^{M}$ from the posterior distribution with density $p(\alpha_r,\beta_r,\theta_p,\boldsymbol{\gamma}\,|\mathbf{Y}^{\text{obs}})$ the quantity in Equation (1) can be approximated with:

$$p(y_{pr}^{\text{rep}} = c|\mathbf{Y}^{\text{obs}}) \approx \frac{1}{M}\sum_{m=1}^{M}\left\{\Phi\left(\gamma_c^{(m)} - \alpha_r^{(m)} - \beta_r^{(m)}\theta_p^{(m)}\right) - \Phi\left(\gamma_{c-1}^{(m)} - \alpha_r^{(m)} - \beta_r^{(m)}\theta_p^{(m)}\right)\right\}$$

for $c = 1, 2, \ldots, C$.

Note that $p(y_{pr}^{\text{rep}} = c|\mathbf{Y}^{\text{obs}})$ is defined for all raters and products—even those rater-product combinations for which no rating was observed.

We are interested in summarizing these posterior predictive probabilities for product p over all raters—even those raters that did not actually rate p. Specifically, we use various methods of graphing

$$\tau_{pc} = \frac{1}{|R|}\sum_{r\in R} p(y_{pr}^{\text{rep}} = c|\mathbf{Y}^{\text{obs}}) \quad (2)$$

for a particular product p and all $c = 1, \ldots, C$. τ_{pc} can be interpreted in two related ways. On the one hand, τ_{pc} is simply the sample average of $p(y_{pr}^{\text{rep}} = c|\mathbf{Y}^{\text{obs}})$ taken over all raters in R. Viewed this way, τ_{pc} is nothing more than a descriptive summary of the collection of $p(y_{pr}^{\text{rep}} = c|\mathbf{Y}^{\text{obs}})$ for $r \in R$. It is also possible to think of a situation where a rater r' is randomly selected from R with equal probability. It is easy to show that the posterior predictive probability that $y_{pr'}^{\text{rep}} = c$ given $\mathbf{Y}^{\text{obs}}$ is τ_{pc}. Thus, τ_{pc} can be thought of as the probability that a randomly chosen rater (from the set of observed raters) will give product p a rating of c given the observed data.

We consider two main ways to plot τ_{pc}. The first approach makes use of a similar visual layout to standard star displays such as those seen in Table 1. The major difference is that rather than using a single color to fill the number of stars equal to the mean rating, our graphical approach color codes each star in a way that reflects the value of τ_{pc} for that product and rating category. This allows the user to directly gauge the probability of a randomly chosen user giving the product in question any rating. We refer to this figure as the "model-based starplot." The second plotting method presents the same information as the model-based starplot but in the form of a barplot where the height of the cth bar for product p is equal to τ_{pc}. We refer to this type of figure as a "model-based barplot."

The advantage of the model-based starplot is that its format is closely related to existing simple starplots commonly used by Web sites. The advantage of the model-based barplot is that it presents the relevant information as vertical distances from a horizontal line, which may be more easily interpretable (Cleveland and McGill 1985; Cleveland 1993). Which method (if either) is to be preferred is clearly an empirical question. Section 4.1 provides results from an experiment we conducted to assess the ability of human subjects to correctly decode information from a variety of graphical displays of the same ratings data. The next section applies our approach to the Mondo Times news outlet ratings.

3.3 An Application: Mondo Times News Outlet Rankings

To illustrate our approach with a real dataset, we examine the Mondo Times data described in Section 2. Figure 1 presents

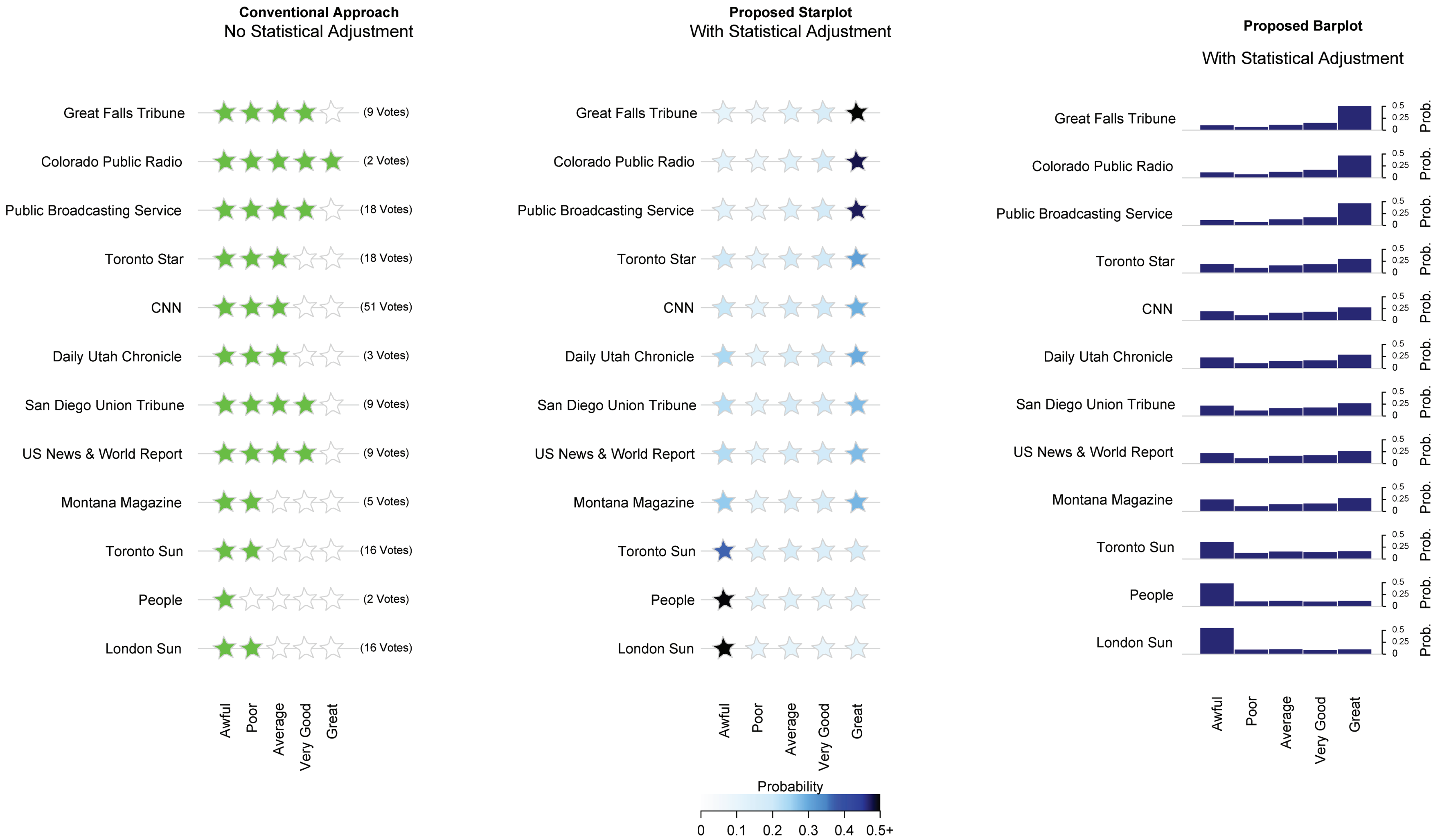


Figure 1. Comparison of conventional approach to presenting ratings data online (left panel) with proposed model-based starplot (center panel) and the proposed model-based barplot (right panel) for the same underlying data. Data are from Mondo Times. Outlets are ordered by decreasing estimated latent content quality.

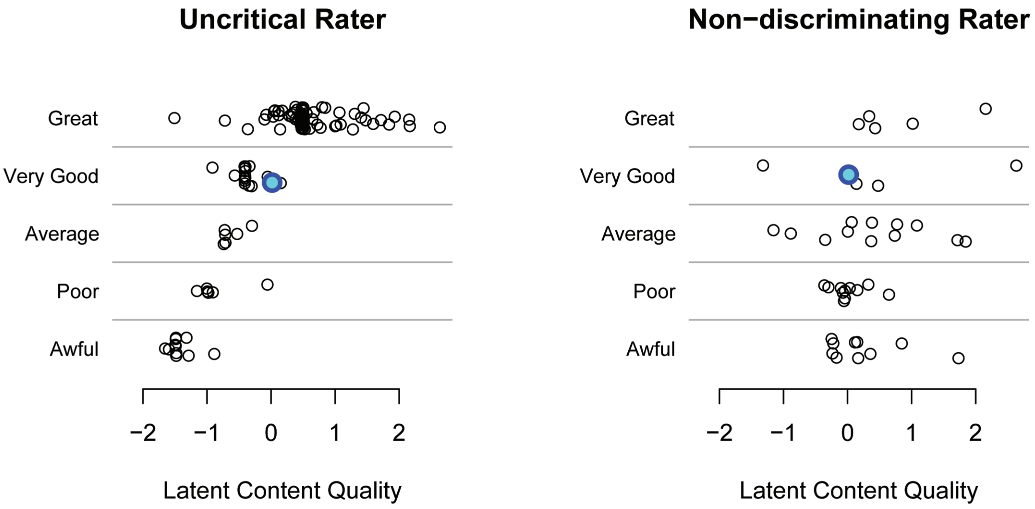


Figure 2. All ratings submitted by the two raters, both of whom rated *U.S. News & World Report* in row 9 of Figure 1. Each panel depicts data from a single rater. Each circle in each panel represents a news outlet rated by the rater in question, randomly jittered for visibility within each of the five rating categories. The filled blue circles represent the rating of *U.S. News*. This figure illustrates why the mean rating of four stars by nine users still leads *U.S. News* to have lower (posterior) probability of a "great" rating than many other outlets with the same number of stars and even the *Daily Utah Chronicle* with a mean rating three stars from only three users. The first panel shows an uncritical user, who rates most outlets as "great", thereby making the "very good" rating for *U.S. News* less meaningful. The second panel shows that one rating of "very good" comes from a nondiscriminating user, thereby providing little information about *U.S. News*.

a comparison of our model-based starplot and model-based barplot to a traditional simple (non-model-based) starplot.

The left panel of Figure 1 provides ratings essentially as they would be presented on the Mondo Web site—without statistical adjustments. We present ratings for only a small subset of outlets to illustrate out presentational approach, not to draw substantive inferences about particular outlets. The number of solid stars represents the rounded mean rating, adjoined by the total number of ratings submitted in parentheses. The center panel depicts the comparable model-based starplot while the right panel depicts the model-based barplot. Differences readily appear. *U.S. News & World Report*, which receives a Mondo rating of four stars, may in fact have *lower* quality content than *PBS* and the *Great Falls Tribune*, each of which also received four stars and *Colorado Public Radio* which was only rated by two raters. Indeed, *U.S. News* may even have lower content quality than the *Toronto Star*, *CNN*, and the *Daily Utah Chronicle*, despite these outlets' *lower* Mondo ratings of three stars. To understand the intuition for the statistical adjustments, Figure 2 plots the observed ratings for two raters of *U.S. News*. Each panel represents all ratings submitted by two raters on the y-axis (randomly jittered for visibility) and the (posterior mean of the) latent content quality as estimated from the IRT model on the x-axis. The first panel shows that *U.S. News* was rated by a noncritical user, who rated more than two thirds of all outlets as "great" (i.e., better than *U.S. News*). If anything, from this user's rating of "very good" we learn that *U.S. News* may be worse than the majority of outlets rated. The second panel plots a user who is largely nondiscriminating, failing to distinguish high and low content outlets in any systematic way compared to the majority of Mondo users. Intuitively, we learn little from such users, as a rating of "very good" does not distinguish the outlet meaningfully. Little information is conveyed by such submissions and our understanding of the quality of *U.S. News*—despite its four-star Mondo rating—remains diffuse.

On the other hand, it is possible for a great deal of information to be provided by a small number of ratings. The second row of Figure 1 depicts the rating information for *Colorado Public Radio*. Only two users rated *CPR*, but both rated it as "great." Standard Web site presentations of the mean rating and the number of ratings might cause users to discount this five-star rating as due to chance variability. However, after accounting for inter-rater differences, the posterior probability that this outlet is of very high quality remains substantial, as indicated by the dark shading of the fifth star in the center panel of Figure 1 and the high probability of a "great" rating in the right panel of Figure 1. Figure 3 provides some intuition for this inference. Here we see that both of the raters in question rated a large number of news outlets *and* both do an excellent job of discriminating between low and high quality outlets. As such, their high marks for this outlet provide considerable information.

Finally, the center and right panels of Figure 1 also illustrate that we have considerable uncertainty about the relative quality of news outlets, particularly in the middle range from rows 4–9 (spanning the *Toronto Star*, *CNN*, the *Daily Utah Chronicle*, the *San Diego Union Tribune*, *U.S. News*, and *Montana Magazine*). The probabilities for these outlets are substantially similar to the prior distribution (i.e., the marginal distribution of ratings without any knowledge about the outlet). This uncertainty stands in sharp contrast to the standard presentation in the left panel, where stars range from 2–4 for those outlets. Despite the large number of ratings submitted for *CNN*, there is considerable heterogeneity within those ratings, leading us to be largely uncer-

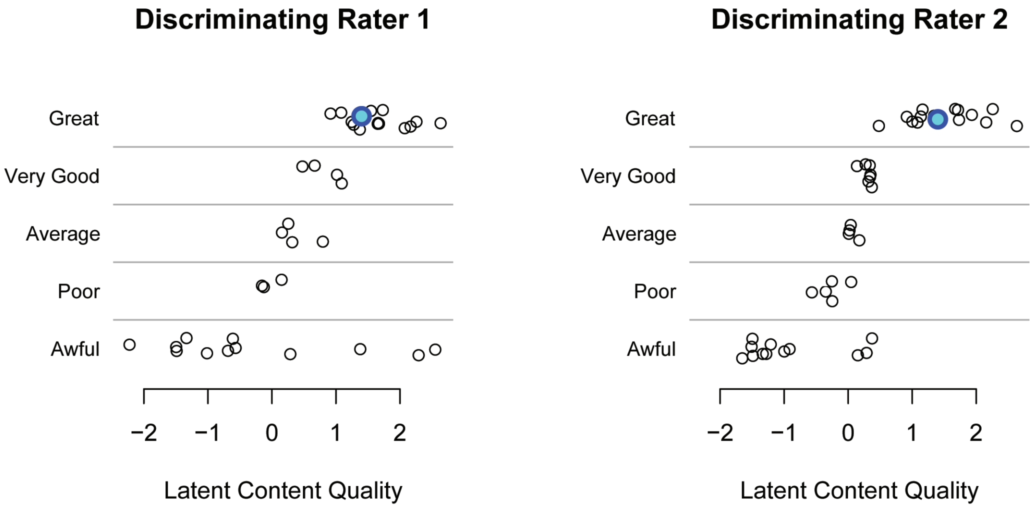


Figure 3. All ratings submitted by the two raters, both of whom rated *Colorado Public Radio* in row 2 of Figure 1. Each panel depicts data from a single rater. Each circle in each panel represents a news outlet rated by the rater in question, randomly jittered for visibility within each of the five rating categories. The filled blue circles represent the rating of *Colorado Public Radio*. This figure illustrates how, even with only two ratings, the (posterior) probability that this news outlet would be rated “great” is quite high. Note that both raters rated many outlets and both raters are very good at distinguishing low and high quality news outlets.

tain as to how a randomly chosen user would rate *CNN*. The statistical adjustment and presentation ensure that users are not left with a false sense of precision as to the quality of outlets.

In short, as is well-known from the IRT literature, statistical uncertainty, rounding errors, and inter-rater differences can have substantial impacts on the interpretation of ratings. Fortunately, meaningful results can be intuitively displayed in our proposed model-based plots.

4. VALIDATING THE APPROACH

Although the theoretical advantages of the approach described above should be clear, there are at least two practical concerns that might be raised. First, one might hypothesize that Internet users will not be able to decode the new figures correctly and thus the figures will be no more effective at conveying accurate information than more standard visual displays. Second, one might be concerned that since model-based figures, by their very nature, rely on modeling assumptions, such figures may end up being much worse than more standard model-free figures in the real world where such assumptions do not hold.

We address both of these concerns here. In Section 4.1 we detail a large experiment we conducted to assess how real Internet users perform at decoding several visual depictions of ratings data. This experiment is similar in spirit to those conducted by Cleveland and McGill (1985, 1987); Wainer et al. (1999), and others. The results strongly suggest that our model-based figures provide more interpretable information than extant approaches. In Section 4.2 we examine how our model-based approach fares against seemingly model-free approaches when the missing data mechanism depends on either the latent product quality, rater attributes, or the ratings themselves. Our results suggest that while our model is not immune to specification bias, it performs considerably better than approaches that simply tally the observed ratings with no statistical adjustments.

4.1 An Experiment in Graphical Perception

To assess how real Internet users decode various visual representations of ratings data we conducted an experiment in which the subjects were all undergraduate students with valid e-mail addresses at Harvard University. The experiment was conducted as follows.

First, we fit the model discussed in Section 3.1 to the Mondo Times data analyzed in Section 3.3. We then calculated the posterior means of all model parameters and generated a new dataset under the assumption that these estimated parameter values are the true parameter values. All missing elements of $\mathbf{Y}$ in the original Mondo Times data remained missing in the simulated data. We then chose eight products from the synthetic data, labeled them products 1 through 8, and devised six graphical displays of the synthetic ratings of these eight products. Each of these six displays represents a treatment arm, each containing the same eight products from the same underlying data. The only aspect varying across the treatments is the visual display.

Treatment 1, displayed in Figure 4, corresponds to the “simple starplot” that is formed by simply taking the rounded mean rating, filling in that number of stars, and then noting the number of ratings of the product in question in parentheses to the right. As noted above, this type of display is widely used to present online ratings data. Treatment 2 is a “simple barplot” that plots the observed fraction of ratings in each of the five possible categories for each product using a barplot. Although such plots are uncommon, some Web sites, such as Amazon, have started to use them in conjunction with the simple starplot. No model-based adjustments are performed in either treatment 1 (the simple starplot) or treatment 2 (the simple barplot).

The remaining four treatments all make use of the same IRT

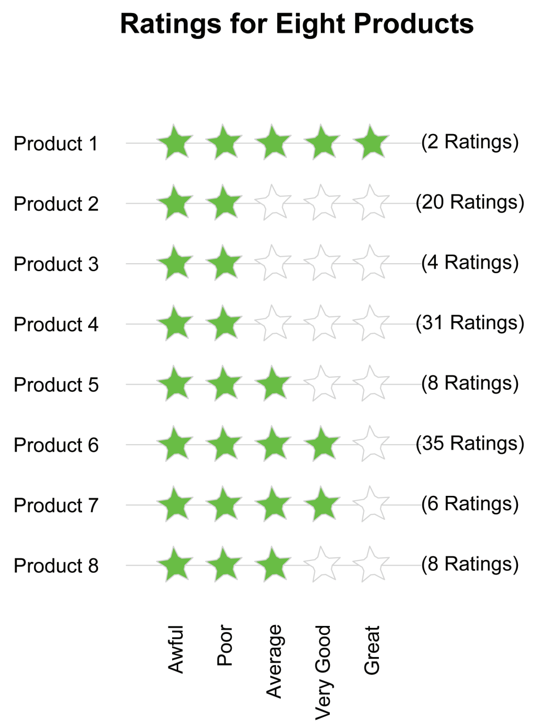


Figure 4. Treatment 1 (simple starplot) from the graphical perception experiment.

adjustment discussed in Section 3.1. All four of these figures plot τ_{pc} for all eight products and all five rating categories. The only differences among these plots is how they display this information. Treatment 3 is a "model-based barplot" in which the height of the cth bar for product p is τ_{pc}. The remaining three treatments are all what we have called "model-based starplots" in that they use the color of the cth star for product p to represent τ_{pc}. Treatment 4, the "model-based starplot (orange)," uses a color scale that is very pale blue from 0 to about 0.2, is then increasingly dark blue from 0.2 to 0.45, and then becomes increasingly orange at points above approximately 0.45. Treatment 5, the "model-based starplot (black)" displayed in Figure 5, is identical to treatment 4 except that orange has been replaced by black. Finally, treatment 6, the "model-based starplot (mono blue)", uses a color scale this is essentially linear in blue.

Subjects were randomly assigned to one of these figures. They were then asked six questions about the products, shown in Table 2 (with correct answers underlined). The correct answer is defined to be the one that has the highest probability of being correct under the posterior distribution based on a correctly specified ordinal IRT model. Using the true parameter values that generated the data to determine the correct answer yields the same correct answers for questions 1–5. For question 6, product 1 has a true quality level most similar to product 5.

The pool of subjects consisted of 6,454 undergraduate students at Harvard University with valid e-mail addresses as of January 1, 2008. Each treatment was randomized to 1/6 of the subject pool. To minimize day-specific response effects, students within each treatment arm were contacted daily over a period of seven days. On any given day, six treatment groups of 153 or 154 students were contacted (6 treatments × 7 days × 153 or 154 = 6454). The overall fraction of students who responded to at least one question was around 0.10. Response rates were comparable across treatments. Further details are available from the authors.

The quantities of interest in this experiment are the probabilities of a subject's answering questions correctly. Operationally, we define this to be the probability that a student who answers at least one of the six questions answers the question of interest correctly. We are interested in how these probabilities vary with treatment assignment. Figure 6 displays posterior mean estimates of these probabilities (based on a binomial model with Jeffreys prior) along with 95% credible intervals.

These results provide very strong evidence that the standard non-model-based figures (the simple starplot and the simple barplot) make it very difficult for users to correctly infer the relative quality of products. For instance, the probability of correctly answering question 1 after seeing either the simple starplot or the simple barplot was less than 0.2. We find similar results for the simple starplot subjects on questions 5 and 6. Subjects exposed to the simple barplot fared slightly better, but were still not as accurate as subjects seeing the model-based figures. The model-based figures that lead to the greatest accuracy are the barplot and the two starplots featuring a nonlinear color scale (model-based starplot (orange) and model-based starplot (black)). Subjects exposed to these figures performed well on all of the questions, with probabilities of correct answers typically well above 0.75 and in some cases very close to 1.0. There is slight evidence that the model-based barplot is easier for subjects to decode than the model-based starplots, al-

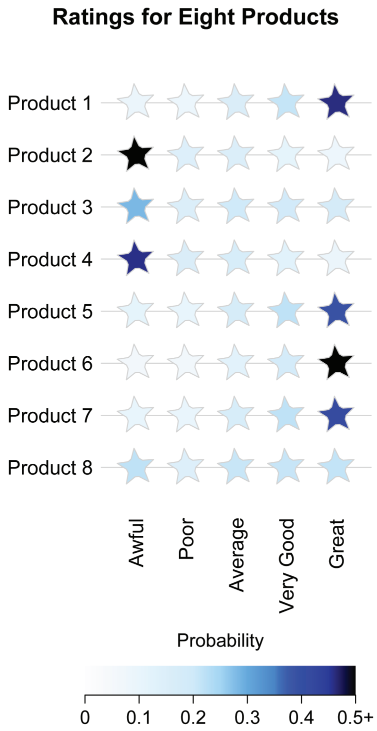


Figure 5. Treatment 5 (model-based starplot (black)) from the graphical perception experiment.

Table 2. Questions asked of all respondents in survey experiment varying graphical display. Correct answers are underlined.

1. Which product do you think is most likely the worst?	1,2, 3, 4, 5, 6, 7, 8, don’t know
2. Which product do you think is most likely the best?	1, 2, 3, 4, 5, 6, 7, 8, don’t know
3. Is 8 likely better than 5?	yes, no, don’t know
4. Is 1 likely better than 6?	yes, no, don’t know
5. Is 3 likely better than 4?	yes, no, don’t know
6. Which product is most similar in quality to product 5?	1, 2, 3, 4, 6, 7, 8, don’t know

though the probabilities of a correct answer from model-based starplot (orange) and model-based starplot (black) are generally quite close to those from the model-based barplot. In general, the experiment strongly suggests that any of these three model-based figures would be a substantial improvement over the simple starplot commonly used on Web sites. Further, it appears that most of the benefit is coming from the use of a model-based adjustment as opposed to a particular plot.

4.2 Exploring Relative Performance With Synthetic Data

To assess how our proposed model-based methods perform when the model is misspecified, we generated ten synthetic datasets under different missing data mechanisms. We then compare our approach relative to the truth as well as simple non-model-based methods. These departures from the basic model could be consistent with either behavioral regularities of real raters or, in some cases, attempts by raters to game the ratings system. All ten synthetic datasets feature 1,000 raters and 500 products. The complete data $\mathbf{Y}$ are generated according to

$$\begin{aligned}\theta_p^{\text{true}} &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \quad \text{for all} \quad p \in P\\ \alpha_r^{\text{true}} &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \quad \text{for all} \quad r \in R\\ \beta_r^{\text{true}} &\overset{\text{iid}}{\sim} \mathcal{N}(1,0.5) \quad \text{for all} \quad r \in R\end{aligned}$$

$$\begin{aligned}y_{pr}^* &= \alpha_r^{\text{true}} + \beta_r^{\text{true}}\theta_p^{\text{true}} + \epsilon_{pr}\\ \epsilon_{pr} &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \qquad \text{for all} \qquad p \in P,\ r \in R\end{aligned}$$

$$y_{pr} = \begin{cases} 1 & \text{if} \quad y_{pr}^* \leq -1.5\\ 2 & \text{if} \quad y_{pr}^* \in (-1.5, -0.5]\\ 3 & \text{if} \quad y_{pr}^* \in (-0.5, 0.5]\\ 4 & \text{if} \quad y_{pr}^* \in (0.5, 1.5]\\ 5 & \text{if} \quad y_{pr}^* > 1.5. \end{cases}$$

Each of the ten synthetic datasets begins with exactly the same complete data dataset. What differentiates the ten synthetic datasets is the missingness mechanism. Here the data were either missing completely at random or the missingness depended on $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, or $\mathbf{Y}$. Full details are available from the authors.

To generate the dataset in simulation 1 we assume:

$$z_{pr} \overset{\text{iid}}{\sim} \text{Bernoulli}(0.9).$$

Thus, approximately 10% of the complete data observations will be included in the simulation 1 dataset and these inclusion probabilities are constant across all rater-product observations. This scenario is an optimistic scenario for model-based analysis, as each rater rates approximately 50 products and each product is rated by approximately 100 raters.

The synthetic dataset of simulation 2 is generated in a similar manner except that the probability of missingness is much higher:

$$z_{pr} \overset{\text{iid}}{\sim} \text{Bernoulli}(0.99).$$

Here, only about 1% of the rater-product observations will be included. The median number of ratings per rater is 5 (5th per-

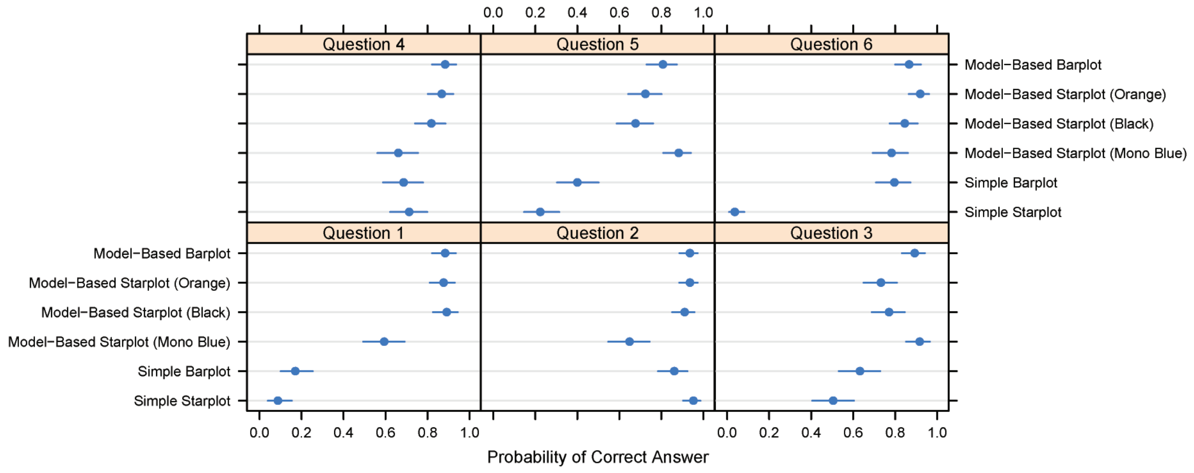


Figure 6. Posterior mean estimates of the probabilities of a correct response to each of six evaluation questions by treatment figure. Dark lines represent central 95% credible intervals. These estimates are based on a binomial model with Jeffreys prior.

centile equal to 2 ratings and 95th percentile equal to 9 ratings) and the median number of ratings per product is 10 (5th percentile equal to 5 ratings and 95 percentile equal to 16 ratings). This degree of missingness is much more realistic than that used in simulation 1.

The synthetic dataset of simulation 3 was generated so that the missingness probabilities depended on θ^{true}:

$$z_{pr} \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - \Phi(-2.5 + 0.3 * |1 + \theta_p^{\text{true}}|)).$$

Here, there is an asymmetric U-shaped relationship between latent product quality and the probability of inclusion, with high and low quality products more likely to be observed than mid-quality products. High-quality products are also more likely to be observed than low-quality products. This represents a situation where raters are more likely to rate good than bad products, and more likely to rate bad than average products. The median number of ratings per rater is 9 (5th percentile equal to 5 ratings and 95th percentile equal to 15 ratings) and the median number of ratings per product is 15 (5th percentile equal to 5 ratings and 95th percentile equal to 50 ratings).

Data for simulation 4 were generated in a similar fashion except that the missingness probabilities depend on α^{true} rather than θ^{true}:

$$z_{pr} \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - \Phi(-2.5 + 0.5 * |\alpha_r^{\text{true}}|)).$$

Again the relationship is U-shaped, although here the inclusion probabilities are symmetric around 0. This mechanism captures a situation where easy-to-please and difficult-to-please raters (high and low α's, respectively) are more likely to rate products than other raters. The median number of ratings per rater is 8 (5th percentile equal to 2 ratings and 95th percentile equal to 29 ratings) and the median number of ratings per product is 21 (5th percentile equal to 15 ratings and 95th percentile equal to 29 ratings).

Missingness for simulations 5–8 was based on the actual values of y_{pr} rather than the model parameters. The missingness indicators for these datasets were generated $z_{pr} \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_{pr})$ where π_{pr} varied with y_{pr}, as displayed in Table 3. Simulation 5 represents a situation where raters who perceive a product as high quality are much more likely to rate that product. Simulation 6 represents a scenario in which raters who perceive a product to be either very bad or very good are more likely to rate that product. Simulation 7 is similar to simulation 6 except that products perceived to be good are somewhat more likely to be rated than products perceived to be bad. Finally, simulation 8 represents a situation in which raters primarily use a subset of the five-point scale (ratings 1, 3, and 5) and products perceived to be good are more likely to be rated.

Simulations 9 and 10 feature extremely low amounts of observed data for each rater and product. In each simulated dataset the number of raters who are observed to rate a given product is 1 plus a Poisson random variable with mean 5. In simulation 9 these raters are chosen from the set of all 1,000 raters with equal probability. In simulation 10 the raters who are observed to rate product p are chosen from the set of all raters with probability proportional to y_{pr} for all $r \in R$. In each simulated dataset the median number of ratings per rater is 3 (5th percentile equal to 1 rating and 95th percentile equal to 6 ratings) and the median number of ratings per product is 6 (5th percentile equal to 3 ratings and 95th percentile equal to 10 ratings).

Table 3. Probability of missingness π_{pr} as a function of y_{pr} for simulated datasets 5 through 8.

	Sim 5	Sim 6	Sim 7	Sim 8
$y_{pr} = 1$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.90$	$\pi_{pr} = 0.95$	$\pi_{pr} = 0.975$
$y_{pr} = 2$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$
$y_{pr} = 3$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.975$
$y_{pr} = 4$	$\pi_{pr} = 0.95$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$	$\pi_{pr} = 0.99$
$y_{pr} = 5$	$\pi_{pr} = 0.90$	$\pi_{pr} = 0.90$	$\pi_{pr} = 0.90$	$\pi_{pr} = 0.95$

Once $\mathbf{Y}^{\text{obs}}$ was generated under each of these ten scenarios, the ordinal IRT model of Section 3.1 was fit to each dataset and quantities of interest were calculated.

First, we are interested in the extent to which the model is able to recover basic information about the true underlying quality of each of the products (e.g., how sensitive the results are to forms of strategic rating). Intuitively, the IRT adjustment removes common forms of strategic rating, such as users rating only several favored outlets at the highest value. To gauge this more systematically, we calculate the Spearman rank order correlation between θ^{true} and the posterior mean of θ. These correlations are quite high—always above 0.8 and typically above 0.9. This is the case even when there is a large fraction of missing data and the missing at random assumption does not hold.

Next we look at how the model-based method performs relative to the simple non-model-based methods that are commonly used. Let

$$\tau_{pc}^{\text{raw}} = \frac{1}{\sum_{r \in R}(1 - z_{pr})} \sum_{\{r\,:\,z_{pr}=0\}} \mathbb{I}(y_{pr}^{\text{obs}} = c)$$

denote the observed fraction of ratings of product p equal to c. This is the measure that is used to construct the non-model-based barplot. Further, let

$$\tau_{pc}^{\text{true}} = \frac{1}{|R|} \sum_{r \in R} \mathbb{I}(y_{pr} = c)$$

denote the true fraction of ratings of product p equal to c in the complete data. This quantity is an in-sample baseline that we would like any method to approach.

With these quantities defined we can define the average percentage reduction in absolute in-sample error for rating category c as:

$$100 \times \frac{1}{|P|} \sum_{p \in P} \frac{\left|\tau_{pc}^{\text{true}} - \tau_{pc}^{\text{raw}}\right| - \left|\tau_{pc}^{\text{true}} - \tau_{pc}\right|}{\left|\tau_{pc}^{\text{true}} - \tau_{pc}^{\text{raw}}\right|},$$

where τ_{pc} is defined by Equation (2). This quantity tells us how much closer to the in-sample ideal the model-based approach comes than the simple unadjusted in-sample fraction. A value of 100 indicates that all of the absolute error in τ_{pc}^{raw} is removed by the model-based approach. A value of 50 indicates that half the absolute in-sample error has been removed.

Looking at the average percentage reduction in absolute in-sample error for all the rating categories in the 10 synthetic datasets we see that the model-based approach almost always removes a large—typically between 20% and 60%—amount of the absolute in-sample error in the simple unadjusted averages. The model-based approach performs relatively well across a range of missing data mechanisms—including those that are not consistent with the assumptions that motivated the model. Additional details are available from the authors.

5. DISCUSSION

In this article we presented and validated simple model-based methods of displaying ratings data—such as are commonly found on Internet sites. Our refinement of basic plots that are already used by the majority of all ratings Web sites, has the promise to better communicate relevant statistical information to everyday users of Web sites, thereby encouraging the sophisticated consumption of goods and services.

That said, there are limitations to our approach. First, our model-based approach dramatically downweights the ratings from users who have only rated a few products. In some applications, such raters will comprise a large fraction of the set of all raters. In situations where there are *not* important inter-rater differences, discarding the ratings of these raters can be detrimental.

Second, although our approach does account for some forms of statistical uncertainty (namely those that arise from sampling variability and rater-specific differences) there are other sources of variability that are not incorporated. The actual missing data mechanism may be quite complicated and failing to correctly model this likely makes the results of our approach falsely precise. Nonetheless, our simulation experiments suggest that our method performs *relatively* well compared to standard non-model-based approaches even when the missing data mechanism is misspecified.

Finally, and perhaps most importantly, the model-fitting approach employed in this article would have to be modified to work efficiently on an industrial scale. Luckily, the sparseness of the underlying ratings data, while unattractive from a purely statistical point of view, may be beneficial for real-time estimation of the model. Since most of the potential ratings are unobserved it will be the case that many of the model parameters will be nearly *a posteriori* independent. This makes it possible to use efficient deterministic approximation algorithms to adjust the model parameters in real-time as new ratings arrive. See Bishop (2008) for an introduction to such ideas and Herbrich et al. (2007) for a large-scale application involving player rankings.

[Received June 2007. Revised August 2008.]

REFERENCES

Albert, J. H. (1992), "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling," *Journal of Educational Statistics*, 17 (Fall), 251–269.

American Association for the Advancement of Science (2004), *Invention And Impact: Building Excellence in Undergraduate Science, Technology, Engineering*, Washington, DC: AAAS, chapter Visualization in Science Education.

Bishop, C. M. (2008), "A New Framework for Machine Learning," in *Lecture Notes in Computer Science LNCS 5050*, New York: Springer, pp. 1–24.

Bock, D. R., and Lieberman, M. (1970), "Fitting a Response Model for *n* Dichotomously Scored Items," *Psychometrika*, 35(June), 179–197.

Boeck, P. De, and Wilson, M. (eds.) (2004), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, New York: Springer.

Bradlow, E. T., and Thomas, N. (1998), "Item Response Models Applied to Data Allowing Examinee Choice," *Journal of Educational and Behavioral Statistics*, 23, 236–243.

Bradlow, E. T., and Zaslavsky, A. M. (1999), "A Hierarchical Latent Variable Model for Ordinal Data from a Customer Satisfaction Survey with 'No Answer' Responses," *Journal of the American Statistical Association*, 94, 43–52.

Cleveland, W. S. (1993), *Visualizing Data*, Murray Hill, NJ: AT&T Bell Laboratories.

Cleveland, W. S., and McGill, R. (1985), "Graphical Perception and Graphical Methods for Analyzing Scientific Data," *Science*, 229(4716), 828–833.

——— (1987), "Graphical Perception: The Visual Decoding of Quantitative Inforation on Graphical Displays of Data," *Journal of the Royal Statistical Society*, Series A, 150, 192–229.

Gelman, A. (2004), "Exploratory Data Analysis for Complex Models" (with discussion), *Journal of Computational and Graphical Statistics*, 13, 755–779.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.

Herbrich, R., Minka, T., and Graepel, T. (2007), "Trueskill^{TM}: A Bayesian Skill Rating System," in *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT, vol. 19, pp. 569–576.

Johnson, V. E., and Albert, J. H. (1999), *Ordinal Data Modeling*, New York: Springer.

Little, R. J.A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley & Sons.

Lord, F. M. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, NJ: Erlbaum.

Meng, X.-L. (1994), "Posterior Predictive *p*-Values," *The Annals of Statistics*, 22, 1142–1160.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002), "The Hierarchical Rater Model for Rated Test Items and Its Application to Large-Scale Educational Assessment Data," *Journal of Educational and Behavioral Statistics*, 27, 341–384.

Quinn, K. M. (2004), "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses," *Political Analysis*, 12, 338–353.

Segall, D. O. (2004), "A Sharing Item Response Theory Model for Computerized Adaptive Testing," *Journal of Educational and Behavioral Statistics*, 29, 439–460.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Wainer, H., Hambleton, R. K., and Meara, K. (1999), "Alternative Displays for Communicating NAEP Results: A Redesign and Validity Study," *Journal of Educational Measurement*, 36, 301–335.