Quant Market Econ (2006) 4: 119–141
DOI 10.1007/s11129-006-3181-x

Is silence golden? An inquiry into the meaning of silence in professional product evaluations

Wagner A. Kamakura · Suman Basuroy · Peter Boatwright

Abstract The world today is rife with product recommendations from professional critics and experts that are available from numerous sources—television, magazines, radio, internet, etc. Very often these recommendations shape our decisions and choices. In this study, we investigate two main issues regarding expert opinions. First, we present an approach that uses information available from every expert, including those who are silent about the product, to obtain a consensus measure of expert opinion. Our model also allows us to obtain a measure of how informative each expert is and how their information content may vary by type of review. More importantly, our overall measure of expert opinion weights the opinion of each expert based on how informative they are at the particular quality level of the product being evaluated. In other words, we provide consumers with a method that reconciles conflicting expert opinions into a summary measure. The second issue we investigate in this paper is the meaning of "silence" in expert opinions. Our model demonstrates that the fact that an expert is silent about a product may imply a positive or a negative review, depending on the expert. We use data from the motion pictures industry to illustrate our approach.

Keywords Expert product recommendations · Expert consensus · Critic silence · Experience goods · Multinomial logit · Latent measurement model

JEL Classification M31

W. A. Kamakura
Ford Motor Company Professor of Global Marketing at Fuqua School of Business, Duke University, Box 90120, Durham, NC 27708
e-mail: kamakura@duke.edu

S. Basuroy
Assistant Professor of Marketing, Florida Atlantic University, Jupiter, FL 33458
e-mail: sumanbasuroy@yahoo.com

P. Boatwright (✉)
Associate Professor of Marketing, Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213
e-mail: pbhb@andrew.cmu.edu

"If you can't say somethin' nice, don't say nothin' at all."
–Thumper, in *Bambi*

Introduction

For experience goods, such as wine or books, experts provide product information that aids consumer choice. At the same time that experts provide details of product characteristics and performance, experts often offer heterogeneous and even conflicting opinions and advice. In the movie industry for instance, Agresti and Winner (1997) studied a small set of well known critics, finding little agreement among them in their "thumbs-up/thumbs-down" opinions. Along with the heterogeneity in critics' evaluations, individual critics offer opinions only on subsets of competing products. The *New York Times* for instance publishes a review of one book each day (Greco, 1997, p. 194). Hence while book reviewers might report on plenty of books, movie critics may remain silent about certain films or even not review most.

While the sparseness of expert opinion in expert-product matrices reflects a reduction in information available about experience goods, the fact that critics are silent on certain products may itself be informative. For instance, critics may be constrained by the number of products they can review. Critics that are lower in the pecking order may be assigned inferior movies to review. Others may select the movies to review in a way that achieves a balanced set of reviews, to avoid being seen as overly negative or overly positive. Or perhaps, some critics may find it easier or more rewarding to write positive reviews.

In this study, we examine the meaning of silence in one population of experts, movie critics. How much information does silence offer? Does it offer information about product quality? Does its meaning depend upon who it is that isn't speaking? At face value, it appears that the public believes the primary information value is in the review content and not in silence, for thumbs up/down is what tends to be reported. We have yet to see a headline, "Ebert did not review _____" or a statistic giving the percentage of critics that reviewed a film. Our results show that critic silence is actually quite informative about movie, a valuable source of relevant information that should not be overlooked. Our research not only studies silence, but it offers a measure of the information content of product reviewers, not just reviews. In our setting, "information content" refers not to the length of a review or number of details covered, rather we refer to the degree to which a review is diagnostic of the product quality. By measuring the degree to which various reviewers are informative about products, our research offers a composite measure that evaluates a product by integrating heterogeneous expert opinions.

As consumers, we currently live in a world that is inundated with product recommendations, ranging from various "experts"—professional critics or real experts (e.g., financial analysts and movie critics) to avid consumers who act as experts (e.g., customer testimonies in amazon.com). Very often these recommendations shape our decisions and choices. For example, Americans routinely seek advice from financial analysts who play a vital role in financial markets. In the entertainment industry, it is reported that more than a third of Americans actively seek the advice of film critics (*Wall Street Journal*, March 25, 1994; B1). About one out of every three filmgoers says she or he chooses films because of favorable reviews. Eliashberg and Shugan (1997) and Basuroy et al. (2003) have shown that movie critics may act both as influencers and predictors and also that critics have significant impacts on box office revenues of movies. For Broadway shows and theaters, Reddy et al. (1998) and Caves (2000) have suggested that theater critics wield "nearly life and death" power over their financial performances.

More generally, such product recommendations are becoming ever more important with the advent of the internet. Recent research on the effects of online recommendation on consumers' online product choices has shown that, in fact, products were selected twice as often if they were recommended (Senecal and Nantel, 2004).

Recommendations or critical reviews are perhaps more important in the case of experience products than search products. Because it is generally difficult or impossible to evaluate experience products prior to purchase, consumers rely more on product recommendations for these products than for search products. King and Balasubramanian (1994) found that consumers evaluating an experience product (e.g., a film-processing service) rely more on expert opinions or recommendations and hybrid decision-making processes than consumers assessing a search product. More recently, in the context of online product choices, Senecal and Nantel (2004) find that recommendations for experience products are significantly more influential than recommendations for search products.

As recommendations and opinions proliferate in the media and the internet, significant concerns have been raised regarding the validity of some recommendations, indicating the need for some objective method of weighting expert opinion. As an example of bias in expert opinion, there is a large body of literature in finance which studies the biases that may be inherent in analysts' stock recommendations and bond ratings. Several of these studies point to inherent conflicts of interest and argue that analysts may favor firms with which their companies have other business dealings (see, for example, Michaely and Womack, 1999). Other studies point out other possible sources of benefits to analysts who provide over-optimistic views (see, Hong and Stein, 1999). Thomson Financial/First Call aggregated analysts' recommendations in July 2001 and found that almost 50% of all recommendations were "buy" while less than 1% were "sell" (Bruce, 2002). Li (2002) analyzed Thompson Financial's I/B/E/S data and found that over a period of 7 years (1994–2000) and out of a quarter million recommendations, the percent of "sell" recommendations consistently average about 2%. In the domain of movies, Basuroy et al. (2003) report that films on average received 43% positive reviews and 31% negative reviews (p. 109). Thus it seems that both in the financial sector as well as in the movies and other experience products, professional analysts and critics are less disposed toward negative reviews.

Recently, Harvey et al. (2000) proposed a two-stage model of advice taking for consumers. In the first stage, consumers assess the diagnosticity of experts from their past performance history. In the second stage, they appropriately utilize the experts' opinions in their own judgment given their respective diagnosticities. Assessment of expert/advisor diagnosticity was found to mediate advisor utilization in subjects' judgments. The framework proposed by Harvey et al. (2000) is an important step in understanding how individuals incorporate advice from multiple experts in their decisions. In their framework, a necessary criterion for advice taking is combining information from various sources in much the same way "cues" are combined in a multiple cue probability learning task. However, rather than the product attribute information being the basis for consumer judgment, expert or advisor opinions, representing an overall assessment of product attribute information, become the key information being integrated to form a judgment. Building on Harvey et al.'s (2000) two-stage model of advice taking, Broniarczyk and West (2002) considered behavioral issues in that context, having examined how consumers' prior beliefs and goals, as well as advisor disagreement, affect consumer assessment and utilization of advisor opinions. Their primary purpose was to examine the extent to which consumers resolve the dilemma of conflicting advice by focusing on a single advisor that has proven predictive ability. In contrast to the extant research, which focuses on how consumers may discard the opinions of some experts in favor of others, our methodology uses information available from every expert,

including those who are silent about the product, to obtain a consensus measure of expert opinion.

Another method of integrating heterogeneous opinion would be to differentially weight the content of all reviews. Some websites on the internet now provide consensus opinions by differentially weighting individual expert's opinions. For example, www.metacritic.com provides measures for movies, video games and music CD's that incorporate inputs from multiple critics. The Metacritic staff collects critical reviews from a selected list of publications and aggregates these reviews into a 0–100 point scale, the Metascore, which is a weighted average of the individual critic scores. Differential weights are assigned to each critic based on the belief that some critics consistently write better (more detailed, more insightful, more articulate) reviews or have more prestige (e.g., *New York Times*, *Variety*) and weight in the industry than others (http://www.metacritic.com/about/scoring.shtml). While this methodology uses differential weights for each expert based on information content in expert opinions, these weights are based on personal, subjective judgment rather than scientific measures of the degree to which each expert is truly informative.

To our knowledge there has been very little research on the assessment of information content in expert opinions. Our research fills this crucial gap in the literature. In this paper, we investigate two main issues regarding expert opinions. First, we provide a methodology that uses information available from every expert, including those who are silent about the product, to obtain a consensus measure of expert opinion. Our model provides the means for interpreting the expert's opinion and allows us to obtain a measure of how informative each expert is. The second issue we investigate in this paper is the meaning of "silence" in expert opinions. There has been virtually no work to our knowledge on the quality implications of "silence" in expert opinions. A majority of films are not reviewed even though the critic might have seen the movie. Our model demonstrates that the fact that an expert is silent about a product may imply a positive or a negative review, depending on the expert.

In the next section we present our measurement model which translates the opinion of heterogeneous experts with sparse reviews into a consensus measure. The model "calibrates" each expert on the basis of their opinion about multiple products, resulting into a response function that translates the expert's opinion into a latent product measure. These response functions provide useful insights into how the experts report their opinions. Once all experts are calibrated based on their past opinions, such weights can be used to integrate their opinions (or lack thereof) about a new product.

Calibrating the opinion of experts

Consider a sample of $i = 1, 2, \ldots I$ experts who review $j = 1, 2, \ldots, J$ products. Instead of assuming an interval or ordinal rating scale, we consider that the experts may rate the products on a nominal scale $y_{ij} = k$, where $k = 1, 2, \ldots, K$ represent multiple, mutually exclusive response alternatives, including the expert's decision not to dispense her opinion on product j. For example, *Variety* magazine reports movie ratings based on critics' opinions that fall into three categories—"negative," "neutral," or "positive."

Based on these observed ratings, our purpose is first to measure the overall (across experts) evaluation of each product and then to calibrate each expert's rating function (with outcomes of positive, negative, neutral, or silent) to the overall evaluation. We define the response function of an expert i as the probability that she gives product j a rating k:

$$P_{ijk} = P(y_{ij} = k|Z_j, \mu_i, \beta_i) = \frac{\exp(\mu_{ik} + \beta_{ik} Z_j)}{\sum_{k'} \exp(\mu_{ik'} + \beta_{ik'} Z_j)} \quad (1)$$

where Z_j = vector of latent variables measuring the acclaim of product j as perceived by all experts; μ_{ik} = intercept for expert i and response category k; β_{ik} = slope coefficient to be estimated for each expert and response category. For identification purposes, the intercepts and slopes of the first response category (μ_{i1}, β_{i1}) are set to zero.

The above response function relates the observed ratings by each expert i to latent, unobserved measures of critical acclaim (Z_j) through a multinomial logit model. Note that acclaim is a latent property of the product, its propensity for acclaim. For some products, perceived product "quality" would be an appropriate label for this latent property. We consider the possibility that acclaim might be multidimensional and that experts may differentially value these multiple dimensions (Broniarczyk and West, 2002). The intercept μ_{ik} measures the expert's propensity to use response category k irrespective of the product being rated. The slope β_{ik} indicates how the odds for response category k by expert i changes as product's latent acclaim increases.

Rather than constraining the response categories to form an interval or even an ordinal scale, the response function in (1) makes no implicit assumptions regarding the nature of the response scale. This makes the model particularly well suited for situations where experts choose not to provide their opinion, allowing us to make inferences regarding the implicit opinions hidden behind the experts' silence. In situations where the experts may provide a neutral, non-committed opinion, the model also will allow us to assess how "neutral" the opinion really is. By inspecting the response functions estimated for each expert, we can also infer whether a "silent" review provides any information regarding acclaim, and whether this "silence" implies a neutral, negative or positive review. If the likelihood of an expert reviewing a product increases (decreases) with product acclaim, silence by this expert implies a negative (positive) opinion. An alternative and elegant specification would be the hierarchical latent variable model for ordinal data proposed by Bradlow and Zaslavsky (1999), which combines a binary probit model for non-response with an ordinal probit for observed responses. The benefit of their formulation is that it considers non-responses as originating from a different process than the subject's opinions; in their application Bradlow and Zaslavsky assume that non-response to an item occurs "if the item is not salient" (Bradlow and Zaslavsky, 1999, p. 46). However, their model uses three latent variables, to account for question saliency, subject's opinion and responsiveness. The benefit of our model is parsimony (using a single latent variable to measure movie acclaim) and simplicity of its estimation. In essence, our model assumes Independence from Irrelevant Alternatives for all response categories, including non-response. This is equivalent to the inclusion of a "no-choice" (or outside good) alternative in a choice model, a common practice in choice-based conjoint analysis. Notice, however, that we only make the IIA assumption conditional on the critic's characteristics and movie quality; our model allows for violations of IIA for each critic (across movies) and each movie (across critics).

While the hierarchical rater model (HRM) proposed by Patz et al. (2002) might seem more general than our simple multinomial response model, we believe the two models have very distinct purposes. The HRM is applied to situations where multiple raters evaluate subjects on multiple items using a graded scale, allowing for the estimation of item as well as rater parameters. In our case, each expert provides only a single "holistic" assessment of each movie. More specifically, the model described above is a special case of the generalized factor model (Kamakura and Wedel, 2001) with multinomial observed variables and standardized normal latent factors.

Along with estimating the parameters of the response functions for all available experts, one must also estimate the latent acclaim of a product. Conditional on the response function parameters, the log-likelihood for multidimensional acclaim Z (the subscript j is ignored to

simplify exposition) is

$$\ell(Z|\{\mu,\beta\}) = \sum_i \left[(\mu_{ik^*} + \beta_{ik^*} Z) - \ln \sum_k \exp(\mu_{ik} + \beta_{ik} Z) \right] \tag{2}$$

where k^* represents the response category used by expert i to rate the product.

With the addition of priors, the model parameters of the two conditional likelihoods $[Z|\mu, \beta]$ and $[\mu, \beta/Z]$ can be estimated in straightforward manner with a Gibbs sampler.

A measure of the informativeness of critics

Once μ_{ik} and β_{ik} have been estimated using available data, these parameters can be used to compute the information function for each critic. Given that the negative Hessian of the conditional likelihood of Z is an information matrix, Eq. (3) provides valuable insights regarding how much information each critic provides about the product's acclaim (Lord and Novick, 1968).

$$\frac{\partial^2 \ell}{\partial^2 Z} = -\sum_i \sum_k \beta_{ik} P_{ik} (\beta_{ik} - \bar{\beta}_i) \tag{3}$$

The information function for an expert can be obtained from (3) as

$$I(Z \mid \mu_i, \beta_i) = \sum_k \left[\beta_{ik} \frac{e^{\mu_{ik} + \beta_{ik} Z}}{\sum_{k'} e^{\mu_{ik'} + \beta_{ik'} Z}} \left(\beta_{ik} - \sum_{k''} \beta_{ik''} \frac{e^{\mu_{ik''} + \beta_{ik''} Z}}{\sum_{k'} e^{\mu_{ik'} + \beta_{ik'} Z}} \right) \right]. \tag{4}$$

The information function is a function of Z, the product acclaim, and it indicates how much expert i contributes to the measurement of overall critical acclaim of a product at the acclaim level Z. This interpretation is similar to the information function of a test item in the educational measurement literature (Bock, 1997), where the information provided by each item in a test varies along the ability continuum. As we will see later in the empirical illustration, each expert provides varying amounts of information about the product's acclaim along the product acclaim space; some experts are more informative about low acclaim products, while others are more informative at the high end.

Interpreting the opinion of movie critics

In this section, we illustrate the features of our approach by applying the multinomial response model to a data set comprised of movies and the critics who reviewed them. Marketing research has shown that movie reviews may have some impact on the box office revenues (Basuroy et al., 2003; Eliashberg and Shugan, 1997). These two studies have shown that the aggregated percentage of positive reviews and the aggregated percentage of negative reviews affect domestic box office revenues. On the other hand, Ravid (1999) did not find any impact of the aggregated percentage of non-negative reviews (i.e., ratio of positive plus mixed reviews over total number of reviews) on domestic box-office revenues. However, a majority of movies do not get reviewed by each movie critic. A movie critic can only write about a fraction of the many movies released each week, even though he/she is likely to

preview most of them. Therefore, the fact that a critic is silent about a particular movie can be caused by multiple reasons. She may have not previewed the movie, either because she was not invited to do so, or because of some preconceptions about it. Or she may have previewed the movie, but decided to write about other movies she had also previewed recently. Many of these multiple reasons for the critic's silence may imply a positive or negative opinion about the movie. Rather than making any prior assumptions regarding the implicit meaning of "silence" by the critic, we will use the multinomial response model described earlier to determine empirically the most likely implied opinion behind each critic's silence.

Data description

The data we use in this illustration is drawn from *Variety* magazine, the leading trade publication in the movie industry. Each weekly issue of *Variety* contains a summary of the opinions expressed by various critics on the movies about to be released in the theaters. For each movie, *Variety* provides the critics' opinions into three categories—"pro" for positive reviews, "con" for negative reviews, and "mixed" for neutral reviews. For our empirical work, the critic's reviews are summarized on a 3-point scale: 1 = negative, 2 = neutral and 3 = positive.

We gathered opinions from 46 top critics (i.e., with the greatest circulation of their primary publication outlets) listed by *Variety* on a sample of 466 movies released between December 1997 and March 2001. This represents 21,436 movie-critic combinations, and covers close to half of all movies reviewed by *Variety* during this period in the "Crix's Picks" section. Table 1 lists the 46 critics in our sample along with a summary of their opinions across all 466 movies. This summary shows that, in general, critics review only a fraction of the movies being released. At one extreme, we have *Petrakis* who reviewed only 23 of the 466 movies. At the opposite extreme, we have *Travers* reviewed 358 of these movies. Consequently, 80% of all movies are reviewed by less than 18 out of the 46 critics in our sample. In other words, the most prevalent data we have is that a critic was silent about a particular movie.

Calibrating the opinion of movie critics

We estimated multinomial response model parameters using our data assuming a uni-dimensional and two-dimensional acclaim space.[1] A multidimensional acclaim space would allow experts to judge the movie along multiple aspects. For example some critics may give a high rating to a movie because of the quality of its cinematography or acting, while others may rate it low because of it lacks general entertaining value (West and Broniarczyk, 1998).). If most critics viewed movies on these entertaining vs. cinematographic quality dimensions, the data would ask for a two-dimensional model.

As for the priors we used for model estimation, our prior for each element of Z was $N(0, 1)$, fulfilling the assumption adopted in the simulated maximum likelihood estimation framework (Kamakura and Wedel, 2001, pp. 518–522). For μ_{ik} and β_{ik} we used independent priors $N(0, 100)$. We estimated each step in the Gibbs routine with Metropolis Hastings samplers, assessing convergence of the chain with Bayesian Output Analysis (version 1.0.1). For μ and β, we used random walk samplers. Our proposal density for $\mu_{ik,t+1}$ was $N(\mu_{ik,t}, 0.2)$, where $\mu_{ik,t+1}$ represents the $(t + 1)$th draw in the chain. For $\beta_{ik,t+1}$ we used $N(\beta_{ik,t+1}, 0.2)$. As for the proposal density for $Z_{j,t+1}$ we used $N(0, 1)$. Of the 100,000 draws in our

[1] In order to eliminate an indeterminacy due to rotation in the two-dimensional model, the loading (β) for the first critic was fixed to zero for the second dimension.

Table 1 Data summary

	Movies reviewed			
Critic	Negative	Neutral	Positive	Silent
Adams	44	26	43	353
Anderson	84	44	68	270
Ansen	43	61	71	291
Bernard	69	43	111	243
Caro	24	30	34	378
Clark	66	43	80	277
Corliss	24	18	24	400
Cunningham	36	36	65	329
Dargis	41	19	32	374
Dumperf	11	8	7	440
Ebert	60	88	182	136
Feeney	2	5	18	441
Gleiberman	64	52	61	289
Gliatto	23	7	19	417
Granger	29	118	103	216
Hoberman	23	36	31	376
Hold en	45	33	43	345
Horwitz	6	9	23	428
Howe	67	27	55	317
Hunter	59	25	57	325
Johnston	61	48	100	257
Kempley	53	27	37	349
Lyons	85	61	196	124
Maltin	23	27	56	360
Maslin	33	40	106	287
Mathews	75	53	90	248
Mitchell	35	18	10	403
Morgenstern	123	71	83	189
O'sullivan	34	31	50	351
Petrakis	8	9	6	443
Rozen	82	83	114	187
Schickel	14	11	27	414
Schwarzbaum	65	57	58	286
Scott	22	13	16	415
Seymour	22	22	32	390
Siegel.J	51	48	109	258
Siskel	10	23	48	385
Stuart	17	9	17	423
Taylor	29	14	31	392
Thomas	7	26	90	343
Travers	106	119	133	108
Turan	55	96	83	232
Van Gelder	9	2	3	452
Wilmington	19	79	93	275
Wilson	6	8	3	449
Wlosczyna	69	37	43	317

chain, we discarded the first 10,000 as burn-in and retained every 20th draw in the remaining sequence.

Based on the Bayesian Information Criterion (40,156 for the uni-dimensional and 40,565 for the two-dimensional model) we came to the conclusion that the 46 critics evaluated the 466 movies along a single dimension. The parameter estimates for the uni-dimensional model are listed in Table 2. The intercepts reflect the general propensity for a critic to use one response category, irrespective of the particular movie being evaluated (recall that intercepts and slopes for the first response category were set to zero for identification purposes). Given that the most prevalent response by any critic is silence, the last intercept for each critic is the largest. For example, *Petrakis*, who reviewed one of the lowest number of movies (second only to *Van Gelder* and *Wilson*), has one of the largest intercepts (second only to *Feeny* who also reviewed very few movies) for the "silence" category. *Lyons*, *Travers* and *Morgenstern*, who express their opinion more often, and write a larger number of reviews, have the smallest intercepts for this category.

The slope coefficients show how the odds of a critic using a response category (relative to the negative category set as the baseline) change with overall critical acclaim. Across most critics (except for *Dumpert*, and *Wilson*), the slope for the positive category is the largest, suggesting that the probability of a positive review increases with movie acclaim. Similarly, the slope for the negative category (set to zero) was the smallest of all categories for most critics, with the exception of *Maltin*. This fact, again, suggests that the probability of a negative review decreases with acclaim, as one would expect, strongly supporting the interpretation of the latent dimension as overall critical acclaim. As for the neutral review, its slope tends to fall between those of the negative and positive response categories, as one would again expect. Interpretation of the slope coefficients for the silent response is more complex, as it varies across critics. For some critics (e.g., *Dumpert*, *Thomas*, *Van Gelder* and *Wilson*) the slope for silent reviews is higher than for positive ones, suggesting that these critics are less likely to review high-acclaim movies.

Note that by modeling the ratings as nominal, relevant information about the structure of the data is not utilized, for there is an intuitive ordering on the effects of positive, neutral, and negative ratings. To overcome this limitation, one could impose sign or order constraints on some of the model parameters, for example by forcing the coefficients for "neutral" to be greater than "negative" and smaller than "positive" (e.g., Bradlow and Zaslavsky, 1999). However, one should avoid continuous linear measures of negative and positive reviews, due to potential asymmetries of these categories. We chose not to impose these constraints to verify empirically their implied order, as another test of face validity for our model.

The relationship between the product's critical acclaim and critics' response probabilities is more easily seen by plotting the response function for each critic, as shown in Fig. 1. The first response function on the top left (*Rozen*) shows the most typical and logical pattern, where the probability of a positive (negative) review increases (decreases) with movie acclaim, and the probability of a neutral or silent review is higher in the middle ranges of acclaim.

The response functions on the top right (*Thomas*) and bottom right (*Petrakis*) show probabilities of positive response that do not increase with product acclaim. On the other hand, the likelihood of a "silent" review increases with acclaim, implying that the absence of a review by *Thomas* or *Petrakis* implies a positive opinion about the movie. The opposite situation occurs with *Corliss* (bottom left of Fig. 1), for whom the probability of a negative review does not change much with product acclaim, while the probability of a "silent" review decreases as product acclaim increases, implying a negative opinion.

The probabilities of negative and positive review are monotonically increasing and decreasing respectively in product acclaim for all critics, as one would expect. Although there

Table 2 Parameter estimates (posterior means) for the multinomial response model

	Intercepts			Slopes		
Critic	Neutral	Positive	Silent	Neutral	Positive	Silent
Adams	−0.38	−0.07	**2.24**	**0.66**	**1.34**	**0.76**
Anderson	**− 0.58**	**−0.54**	**1.25**	**0.49**	**1.46**	**0.47**
Ansen	**0.34**	−0.73	**1.95**	**0.73**	**2.47**	0.17
Bernard	−0.19	**0.46**	**1.56**	**1.07**	**1.83**	**0.92**
Caro	**0.58**	0.27	**3.14**	**1.01**	**2.12**	**1.13**
Clark	**−0.52**	−0.51	**1.47**	**0.95**	**1.86**	0.16
Corliss	−0.43	−0.54	**2.82**	**0.58**	**1.21**	0.09
Cunningham	0.05	**0.27**	**2.31**	**0.79**	**1.55**	**0.49**
Dargis	**−0.77**	**−1.21**	**2.24**	0.15	**1.78**	0.14
Dumpert	−0.16	−0.31	**3.94**	**0.76**	**0.72**	**1.01**
Ebert	**0.88**	**1.43**	**1.42**	**0.81**	**2.33**	**1.56**
Feeney	−0.29	0.77	**4.74**	−0.85	**1.53**	0.23
Gleiberman	−0.06	−0.33	**1.69**	**0.93**	**1.80**	**0.73**
Gliatto	**−1.21**	−0.62	**2.91**	0.06	**1.10**	0.13
Granger	**1.58**	**1.12**	**2.19**	**0.69**	**1.71**	**0.75**
Hoberman	0.19	−0.76	**2.80**	0.57	**1.51**	−0.32
Holden	−0.16	0.03	**2.20**	**0.54**	**1.13**	**0.84**
Horwitz	−0.01	**1.09**	**4.21**	−0.05	**1.34**	**0.83**
Howe	**−0.78**	−0.53	**1.69**	**0.59**	**1.68**	**0.64**
Hunter	**−1.00**	−0.25	**1.73**	**0.81**	**0.95**	0.21
Johnston	−0.12	0.18	**1.59**	**0.89**	**1.78**	**0.52**
Kempley	**−0.56**	−0.49	**2.02**	**0.70**	**1.37**	**0.68**
Lyons	−0.30	**0.83**	**0.42**	0.24	**0.64**	**0.33**
Maltin	0.11	0.10	**2.82**	0.00	**1.07**	**−0.65**
Maslin	0.26	**0.65**	**2.24**	**0.68**	**1.88**	**0.35**
Mathews	−0.09	0.08	**1.50**	**1.21**	**1.97**	**0.95**
Mitchell	−0.56	**−1.40**	**2.57**	**0.67**	**1.35**	**0.67**
Morgenstern	**−0.71**	**−0.73**	**0.46**	**1.07**	**1.34**	0.10
O'sullivan	0.06	0.20	**2.52**	**0.47**	**1.68**	**0.80**
Petrakis	0.22	−0.03	**4.40**	**0.51**	**1.38**	**1.27**
Rozen	**0.35**	**0.22**	**1.19**	**1.33**	**2.24**	**0.97**
Schickel	−0.43	0.14	**3.35**	0.63	**1.25**	0.20
Schwarzbaum	0.09	−0.44	**1.73**	**1.01**	**2.06**	**0.86**
Scott	−0.47	−0.51	**3.04**	0.37	**1.32**	**0.63**
Seymour	0.19	**0.38**	**3.10**	**0.72**	**1.51**	**0.90**
Siegel.J	0.00	**0.55**	**1.74**	**1.00**	**1.57**	**0.46**
Siskel	**0.97**	**1.61**	**3.83**	**0.98**	**1.53**	**1.08**
Stuart	−0.58	−0.07	**3.32**	0.52	**1.14**	**0.66**
Taylor	**−0.98**	−0.67	**2.62**	**0.78**	**1.40**	−0.02
Thomas	**1.20**	**2.62**	**3.96**	**0.50**	**1.02**	**1.26**
Travers	**0.75**	−0.59	**0.60**	**1.85**	**4.29**	**1.12**
Turan	**0.55**	**−0.79**	**1.44**	**0.73**	**2.46**	0.01
Van Gelder	−1.17	−0.82	**4.33**	0.64	0.76	**1.16**
Wilmington	**1.48**	**1.27**	**2.74**	**0.50**	**1.62**	**0.57**
Wilson	0.15	−0.87	**4.26**	0.31	0.11	**0.49**
Wlosczyna	−0.27	−0.53	**1.87**	**0.86**	**2.07**	**1.22**

Note: Estimates in bold are more than 1.96 posterior standard deviations away from zero

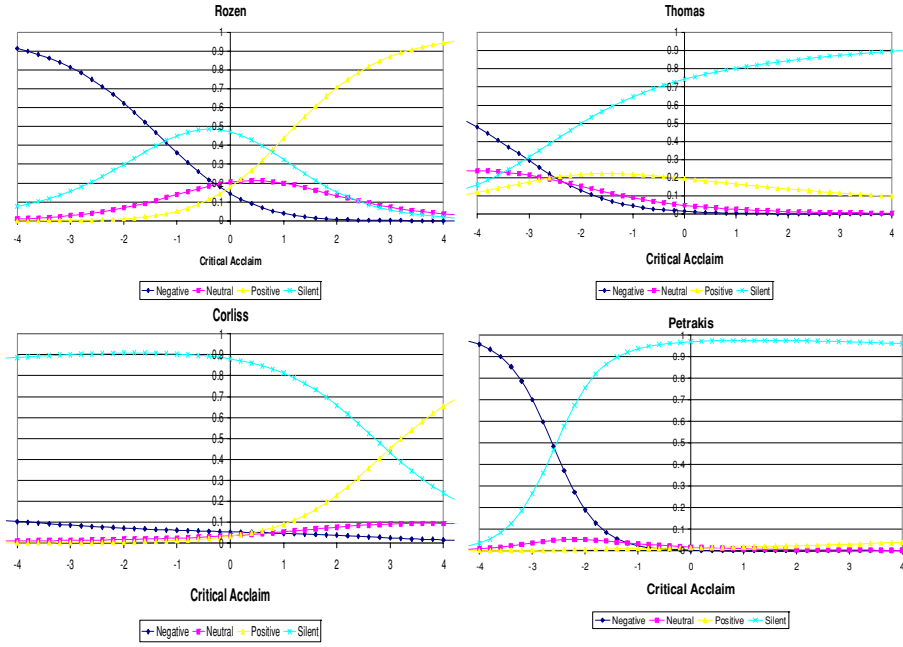


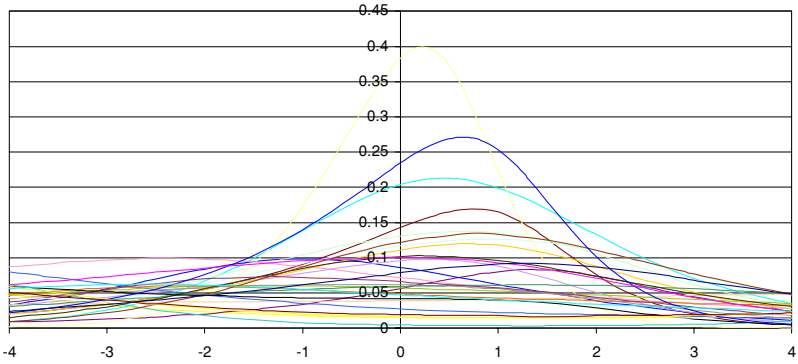
Fig. 1 Response functions for selected movie critics

is a general trend across most critics for the neutral evaluation to be somewhere in between negative and positive, there is more variation across critics for this response category than for the negative and positive categories. As shown in Fig. 2(a), for many of the critics the likelihood of a neutral review indeed peaks at average movie acclaim, as one would expect. However, Fig. 2(b) shows that for 6 of the 46 critics, the likelihood of a neutral review is higher for low-acclaim movies (keeping in mind that quality is measured in a standardized normal scale), implying a negative opinion. Figure 2(c) shows that for four other critics the likelihood of a neutral review increases with movie acclaim, implying a positive opinion.

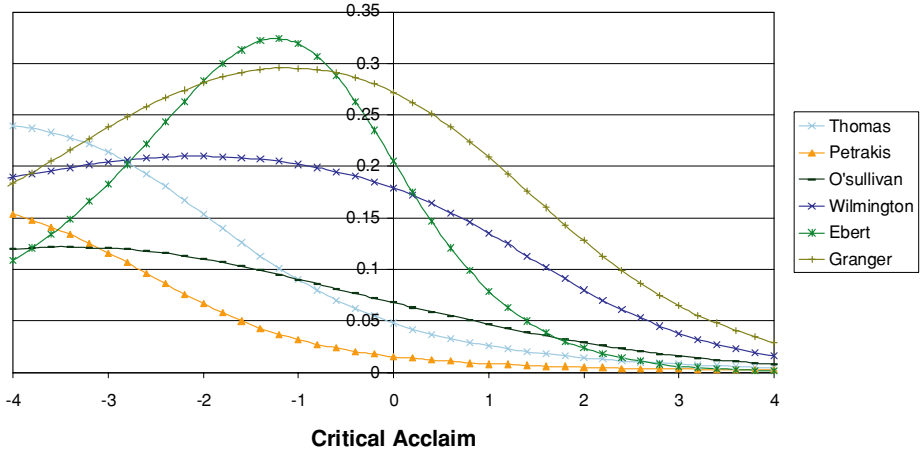
Figure 3(a) shows that 13 of the 46 critics studied are more likely to be silent when the movie is of low acclaim than when it is of high acclaim. This implies that silence by these reviewers is a clue that the movie might be of lower acclaim. Figure 3(b) shows the opposite for 3 other critics, for whom silence implies a positive opinion.

Measuring movie quality from critic reviews

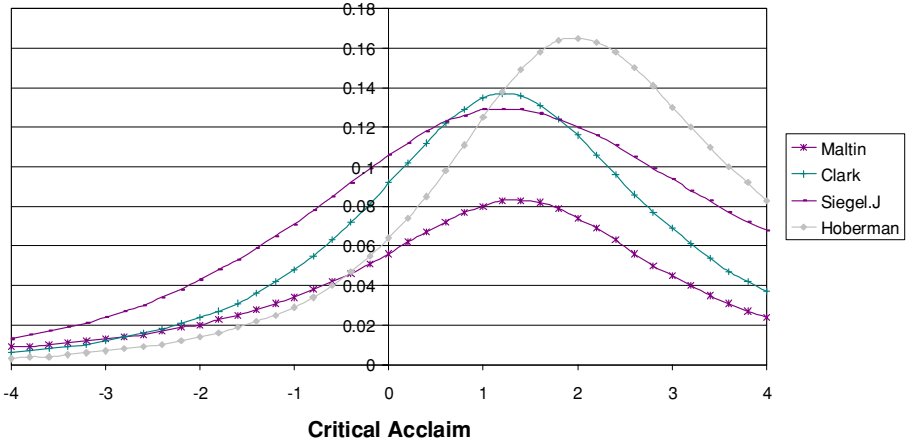
As discussed earlier, a summary measure of critical acclaim for a particular movie based on critic review is the information function $I(Z|\mu_i, \beta_i)$, shown in Eq. (4). We estimated the posterior distribution of each critic's information functionby computing it for all draws of μ_i and β_i over a grid of Z values. We report the (posterior) mean of these computations for each value of Z for each critic. As the information functions reveal, some critics are more informative than others. Moreover, a critic is not uniformly informative over the whole range of the quality space. Therefore, once the response function for each critic is known, one can selectively seek opinions, depending on how the information on the movie's acclaim will be used.



(a)

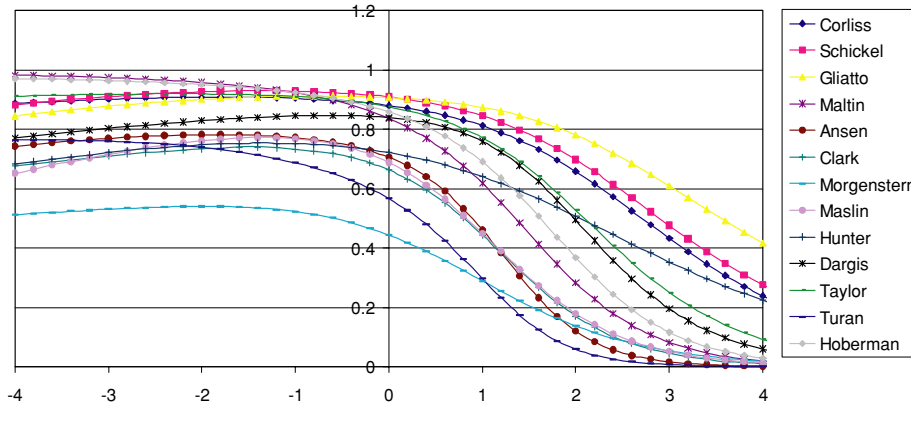


(b)

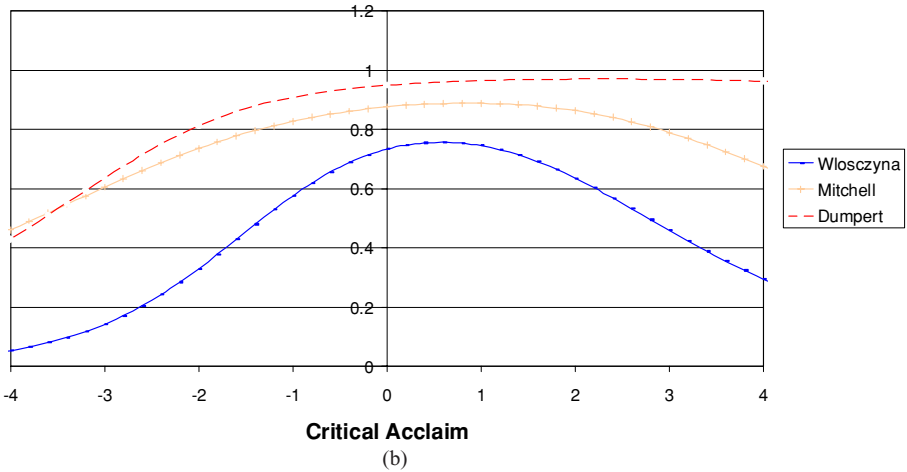


(c)

Fig. 2 (a) Response probability for neutral reviews, (b) Response probability for neutral reviews implying low critical acclaim and (c) Response probability for neutral reviews implying high critical acclaim



(a)



(b)

Fig. 3 (a) Response probability for silent reviews implying low critical acclaim and (b) Response probability for silent reviews implying high critical acclaim

Figures 4(a) and (b) show the information functions for the most informative and the second most informative groups of critics, respectively. Based on Fig. 4(a), *Travers*, *Turan* and *Ansen* are three critics who provide the most information regarding movie acclaim. However, *Turan* and *Ansen* are most informative for high-acclaim movies, providing limited information in the low-acclaim range. *Travers*, on the other hand, provides information on a broader acclaim range.

A consumer wishing to avoid low-acclaim movies should give priority to the opinions by *Ebert*, *Mathews*, *Caro*, and *Wlosczyna*, who provide the most information in the below-average acclaim range. A consumer with a limited budget of time or cash trying to focus only on the best movies should seek the opinion of *Turan*, *Ansen*, *Dargis*, *Hoberman* and *Clark* who provide the most information at the high end.

Given that experts differ on how informative their opinions are, and on what range of product acclaim they are most informative, it would seem intuitive to weight their opinions

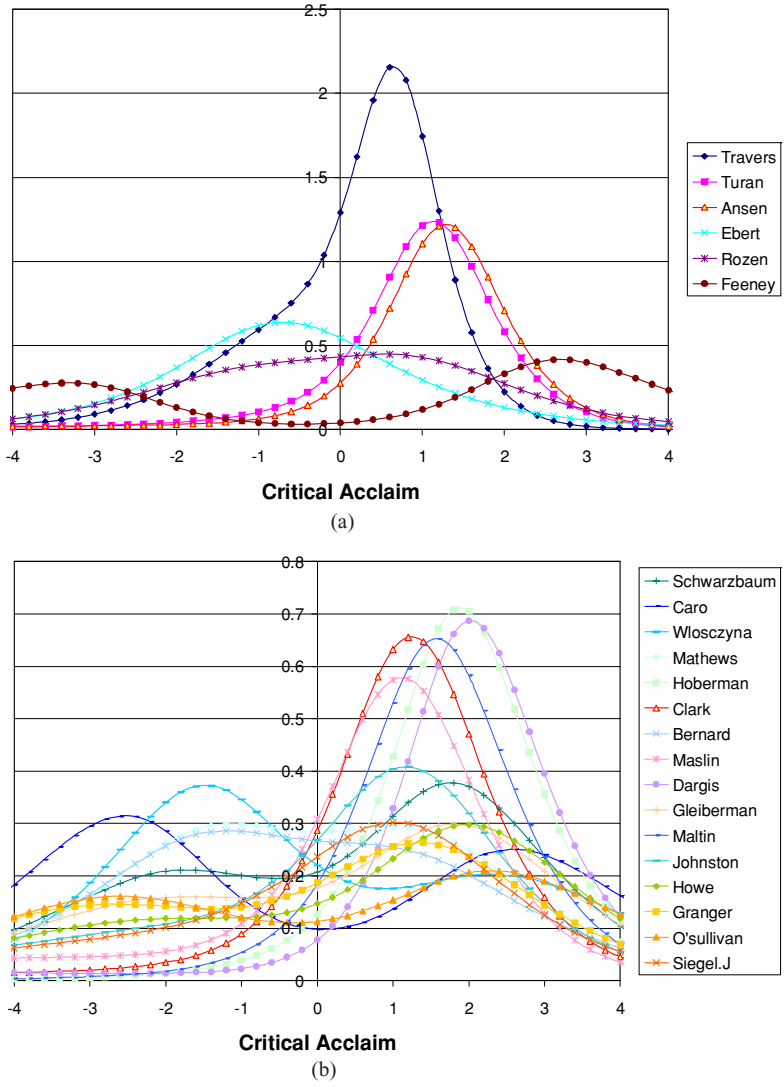


Fig. 4 (a) Information function for the most informative critics and (b)Information function for the next group of informative critics

differentially when forming an overall consensus measure of product acclaim. This is essentially what the nominal response model does, through the estimation of product acclaim as shown in (1)–(3). Rather than simply aggregating all opinions regarding a product into a simple score, the multinomial response model will weight each expert's opinion according to how informative the expert is at the product's acclaim level.

Given that we already have the opinion of all 46 critics on all 466 movies, we obtained a measure of movie acclaim (Z) for each movie using all critics. The 40 best and worst movies based on this measure are listed in Table 3. The reader is invited to compare this ranking with her own opinion about the movies, as a measure of the reader's preferences with those of

Table 3 Top and Bottom 40 based on Critical Acclaim

Top 40	Score	Bottom 40	Score
Being John Malkovich	2.013	Lost Souls	−2.044
American Beauty	1.989	I Still Know What You Did Last Summer	−1.886
The Insider	1.894	Mercury Rising	−1.820
Bowfinger	1.774	Battlefield Earth	−1.768
The Winslow Boy	1.741	3000 Miles to Graceland	−1.703
Election	1.690	Beautiful	−1.703
In the Mood for Love	1.662	Saving Silverman	−1.669
The Truman Show	1.607	Sweet November	−1.616
Shakespeare in Love	1.603	Desperate Measures	−1.611
Bulworth	1.599	Dungeons and Dragons	−1.588
A Simple Plan	1.577	Instinct	−1.569
The Iron Giant	1.576	Red Planet	−1.555
Traffic	1.540	The Watcher	−1.505
Erin Brokovich	1.535	Whipped	−1.488
Almost Famous	1.522	Monkeybone	−1.476
Beloved	1.495	The Next Best Thing	−1.470
Girlfight	1.474	Flintstones in Viva Rock Veg	−1.470
Thirteen Days	1.470	Baby Geniuses	−1.465
Waking Ned Devine	1.446	Double Take	−1.451
The Mask of Zorro	1.430	Lost and Found	−1.451
Tarzan	1.430	One Tough Cop	−1.437
Chicken Run	1.426	The Wedding Planner	−1.428
Toy Story 2	1.421	Head Over Heels	−1.407
Gladiator	1.418	Lost in Space	−1.405
Three Kings	1.413	I'll Be Home For Christmas	−1.395
The Straight Story	1.408	Drop Dead Gorgeous	−1.389
Primary Colors	1.404	Deuce Bigalow: Male Gigolo	−1.386
Saving Private Ryan	1.402	Jack Frost	−1.385
Antz	1.393	Drowning Mona	−1.378
All About My Mother	1.390	My Favorite Martian	−1.351
Meet the Parents	1.390	Inspector Gadget	−1.349
The General	1.387	Bless the Child	−1.347
Titanic	1.373	The Mod Squad	−1.321
Space Cowboys	1.363	The Replacements	−1.297
Memento	1.350	Company Men	−1.295
Crouching Tiger Hidden Dragon	1.343	The Skulls	−1.294
The Celebration	1.341	See Spot Run	−1.251
There's Something About Mary	1.335	Hollow Man	−1.236
High Fidelity	1.334	The King and I	−1.231
Boys Don't Cry	1.332	The Replacement Killers	−1.230

the critics. We note that these measures of critical acclaim are based on limited information (opinion of 46 critics, who do not necessarily agree with each other), and therefore the measures have a measurement error associated with them. The standard deviation of the posterior distribution is plotted in Fig. 5, against the estimate of movie quality for each of the 466 movies reviewed. Figure 5 indicates that the measurement error tends to decline with critical acclaim, which is consistent with the fact that most critics are more informative at the positive end of the standardized acclaim continuum, as we saw earlier on Fig. 4.

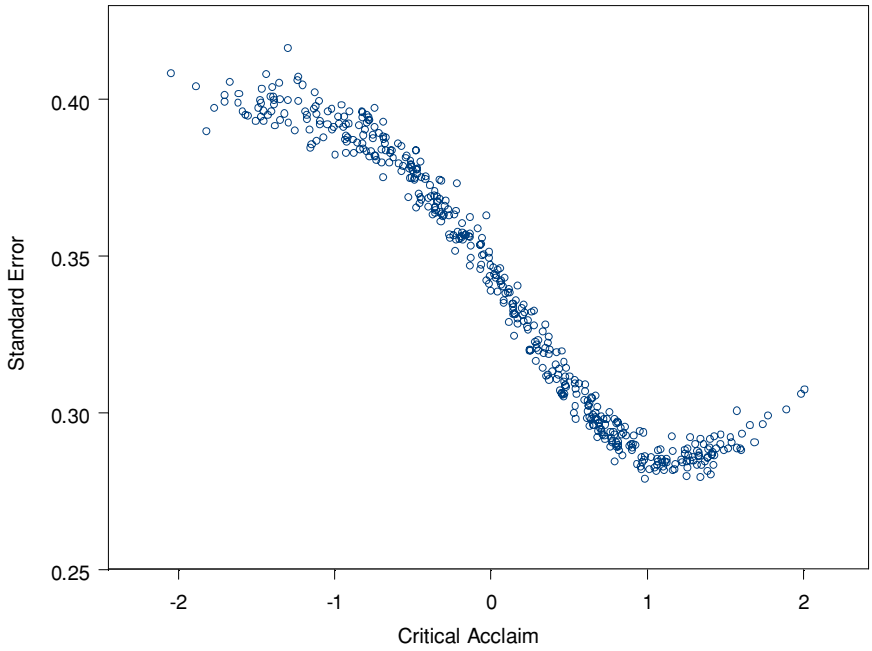
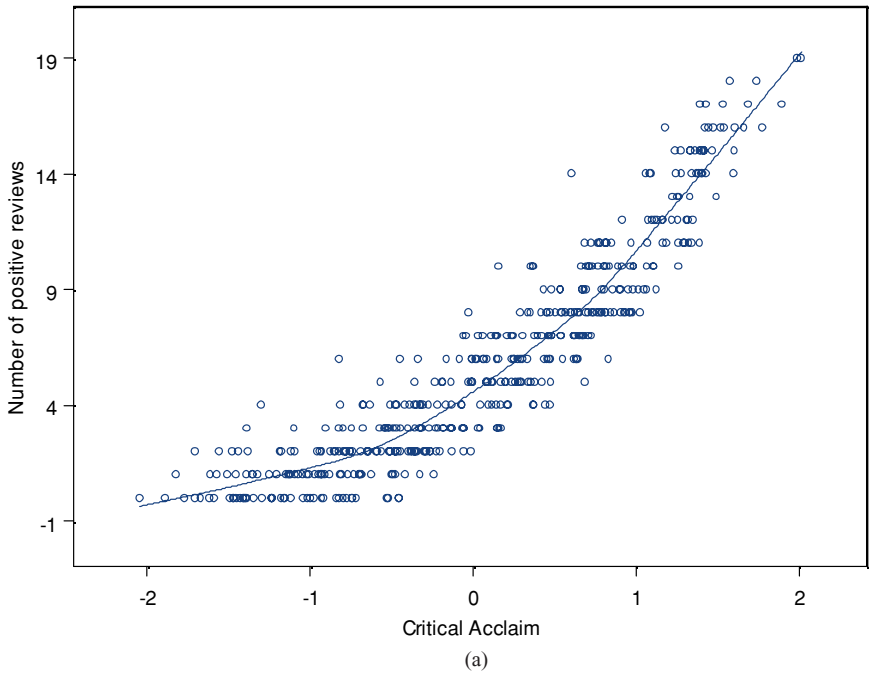


Fig. 5 Standard error of measurement for critical acclaim

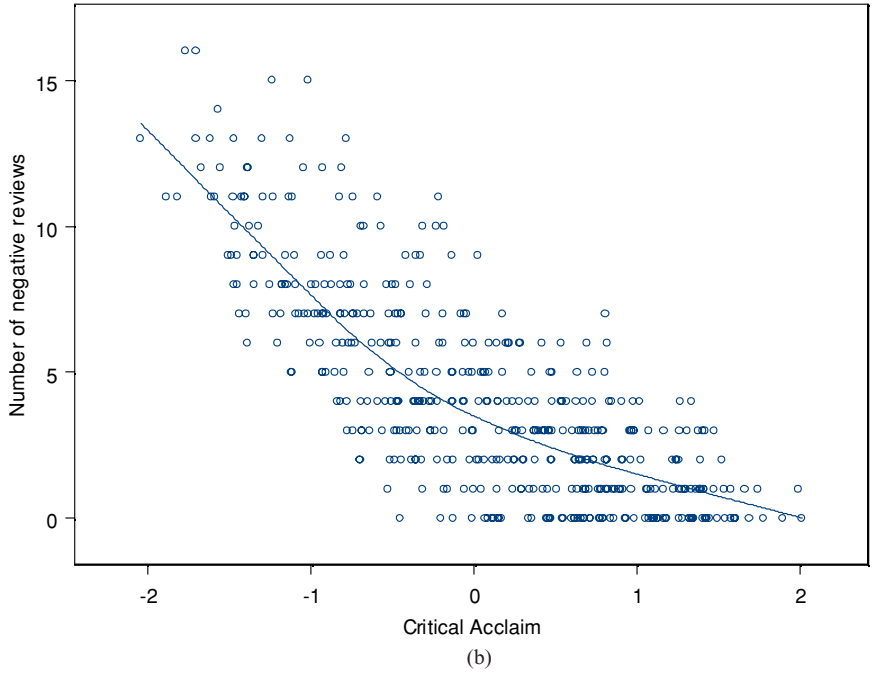
In order to check the validity of our measure of movie acclaim we compare these estimates with the raw data. Figures 6(a) and (b) compare our measurements of movie acclaim with the number of positive and negative reviews obtained by each movie, showing a clear relationship between the raw data and our measure of movie acclaim, as one should expect, given that movie acclaim was estimated on the basis of these raw ratings, and all critics showed a clearly monotonic relationship between movie acclaim and the likelihood of positive or negative reviews.

Figure 6(c) makes the same comparison, but now with the number of neutral reviews attained by each movie. Here we can also see a valid relationship between movie acclaim and the raw data, with neutral reviews happening more often at the mid-range of movie acclaim. Figure 6(d) shows a more tenuous relationship between the number of "silent" reviews and our measure of movie acclaim, suggesting a slightly larger average of silent reviews in the mid-range of movie acclaim. This weaker relationship reflects the fact that 8 critics are more likely to be silent when the movie is of poor quality while 3 others tend to be silent when the movie is of high acclaim.

Additionally, we compare our measurement of movie acclaim, based on critic reviews, to the box-office performance of the movies. More specifically we look at the total ticket sales in the opening week. Rather than a direct validity test for our measurement model, this comparison allows us to verify whether market performance in the opening week has any relation to critical acclaim. As shown in Fig. 7(a), the answer is "no"; there is little relationship between critical acclaim and ticket sales. One then wonders whether this weak relationship between movie acclaim and ticket sales is due to any distortion or bias in our measure of movie acclaim from critic reviews. We verify this by comparing ticket sales to a measure of acclaim that is not related to the measurement model, coming directly from the

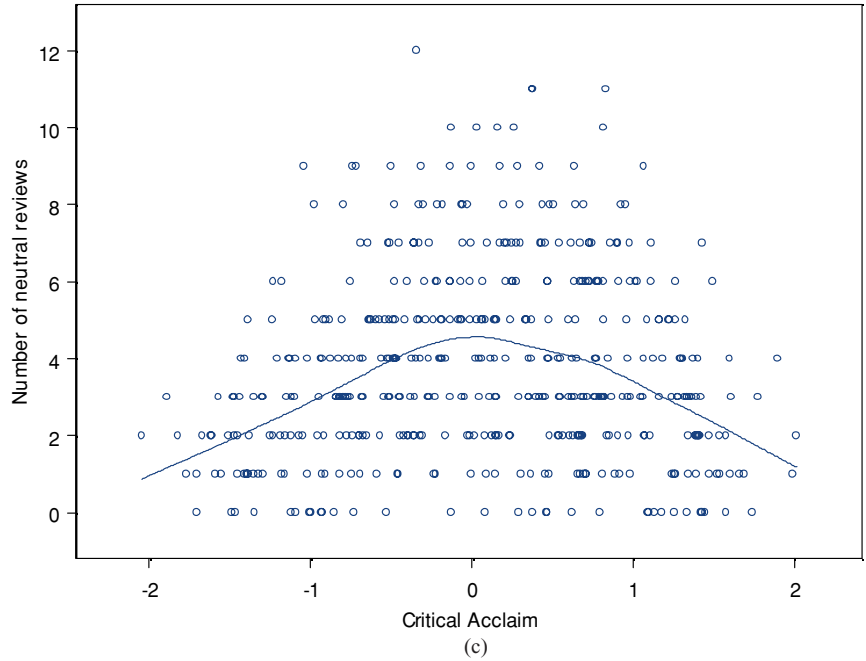


(a)

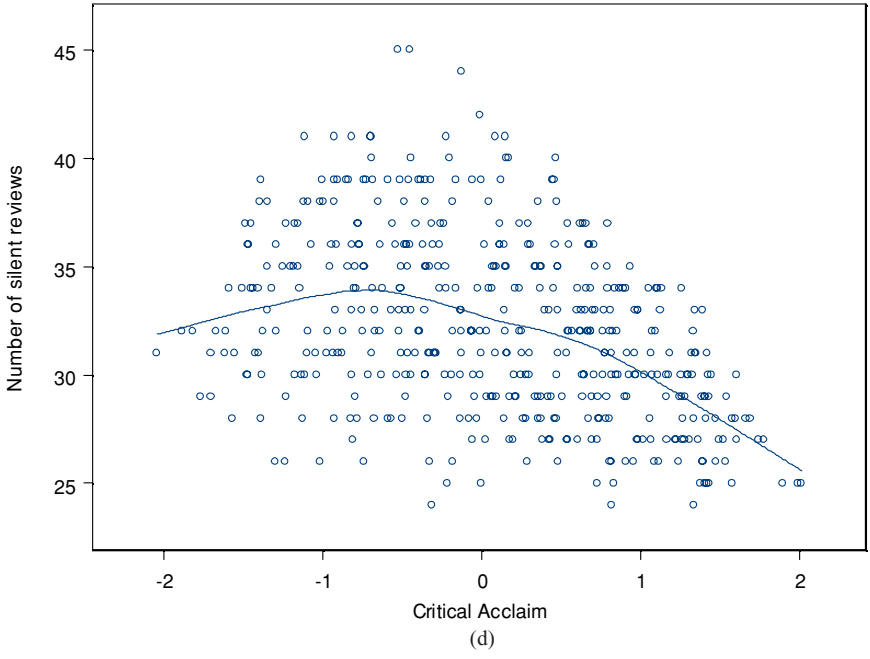


(b)

Fig. 6 (a) Critical Acclaim and the number of positive reviews, (b) Critical Acclaim and the number of negative reviews, (c) Critical Acclaim and the number of neutral reviews and (d) Critical Acclaim and the number of silent reviews (*Continued on next page*)



(c)



(d)

Fig. 6 (Continued)

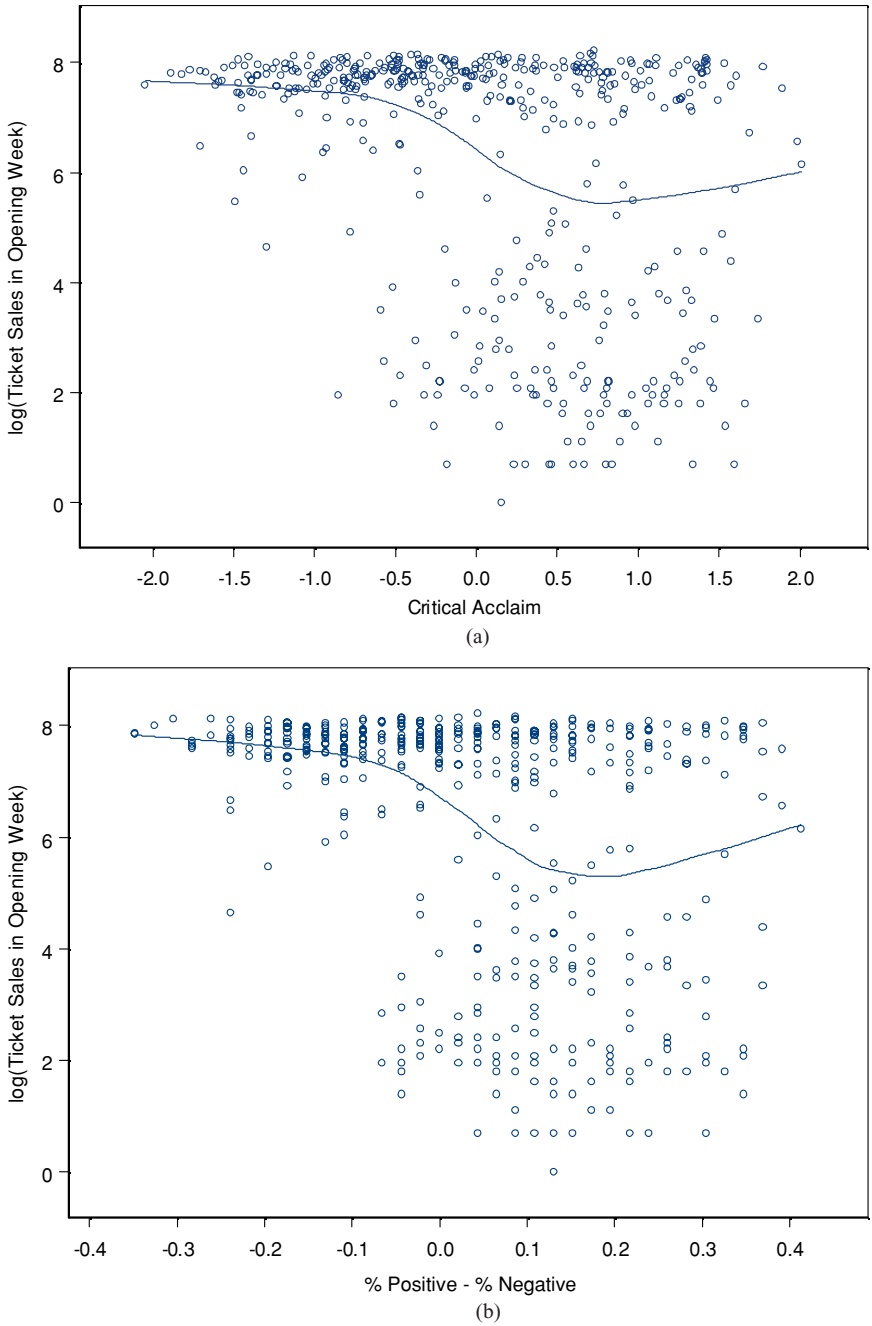


Fig. 7 (a) Comparison of movie acclaim and ticket sales in opening week and (b) Comparison of raw ratings (% positive - % negative) and ticket sales in opening week

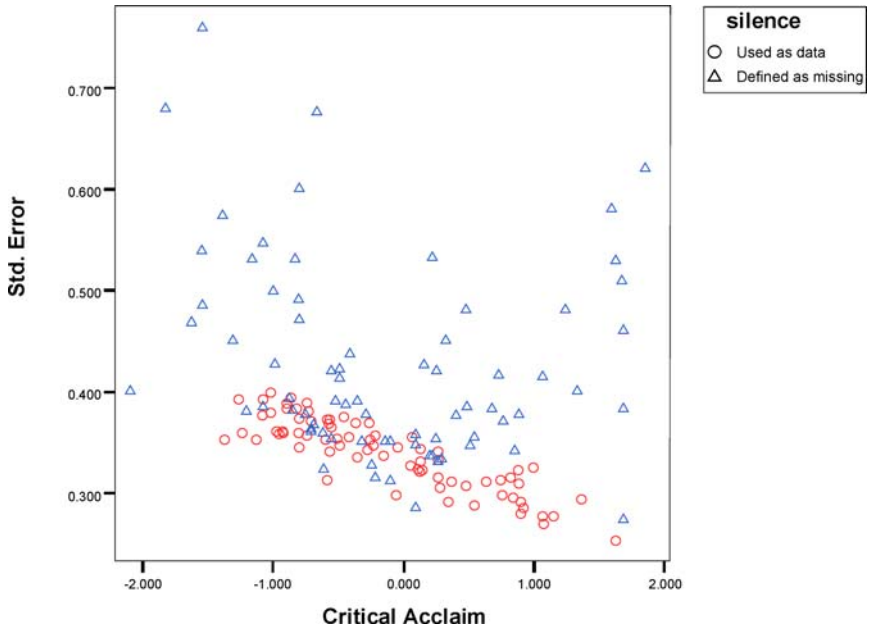


Fig. 8 Acclaim measurements and their standard errors on a holdout set of 75 movies

raw data: the difference between the proportion of positive and negative reviews across the 46 critics. This comparison (Fig. 7(b)) also shows little relation between critical acclaim and market performance in the opening week, suggesting that popularity of a movie has little to do with the consensus among experts. Most importantly, this lack of relationship is observed in the raw data and therefore is not a consequence of our measurement of acclaim.

Measuring the information content of "silent" reviews

Our previous analyses of the response functions for each of the 46 critics suggest that "silent" reviews by some critics provide some implicit information about movie acclaim. In order assess the precision added via information from "silent" reviews, we used the critics' opinions on a holdout set of 75 movies reviewed immediately after our calibration period to obtain two measures of movie acclaim. Both measures were obtained from the multinomial response model (Eq. (2)), except that one assumed "silent" reviews to be missing data, using only the published reviews, while the other used the calibrated model discussed earlier, taking silence as another response category. Despite the fact that the 75 movies received only 25.1% of a total of 3450 possible reviews, the two measures of movie acclaim were highly correlated ($r = 0.95$). However, as shown in Fig. 8, which displays the estimates of movie acclaim from the two models against their standard errors, the model that takes "silent" reviews as data produces lower measurement errors, confirming that these "silent" reviews provide information about movie acclaim.

Figure 8 also illustrates another intriguing aspect of critical reviews. Not only is critic silence informative, but silence is most informative for the better movies, in that standard errors are lowest for the most highly acclaimed movies. If we ignore silence, quality assessments are

least precise for the best and the worst movies, for standard errors are largest for the acclaim extremes. But for the acclaim measure that includes silence, precision is higher everywhere (standard errors are lower), but precision is highest for the more highly acclaimed movies. To the extent that one wants quality assessments of the best movies, a top 10 list or top 100 list, silence is golden.

Conclusions

Product recommendations from professional critics and experts are ubiquitous, proliferating across all possible media—television, magazines, radio, internet, etc. Such recommendations often shape the decisions and choices of time-constrained modern consumers. There has been very little research on the assessment of information content in expert opinions. In this paper, we investigate two main issues regarding expert opinions. First, we provide a methodology that uses information available from every expert, including when they are silent about the product, to obtain a consensus measure of a product's critical acclaim. The approach outlined in this paper extends beyond movies to other product categories. The method can be applied to any situation where multiple experts provide opinions on the same products/services—a common phenomenon in the financial products, restaurants, theaters, books, etc. For instance, restaurant critics from the *New York Times* routinely rate restaurants on a five-star system.

Our model allows this consensus measure to be multi-dimensional, where for movies, books, and plays the acclaim may include dimensions such as entertainment value, acting, or complexity of the plot. In our empirical movie data, the consensus measure turned out to be uni-dimensional, a result that we find interesting by itself. With the complexity of movies and the diversity of opinions and viewpoints of critics, it is not clear *a priori* that a uni-dimensional movie acclaim construct would arise.

In addition to estimating a multivariate consensus of acclaim, our model provides ways for interpreting the expert's opinion, and the model allows us to obtain a measure of how informative each expert is. The generality of our model allows for a variety of different formats of critic viewpoints, not simply a yes/no outcome. For instance, opinion categories could be qualitative such as "dramatic", "riveting", "humorous", and so on, as long as these response categories are common across experts and products being evaluated. Not only will the model identify which opinion categories are associated with underlying latent acclaim factors, but the methods also reveal the contribution of each expert to the overall critical acclaim measures. In our empirical example using movies, the results did show that some critics are more informative than others. Possibly more importantly, our results showed that critics are not uniformly informative for movies of different levels of acclaim.

A consumer can use these results of our model in a practical way. For example, our results indicate that if a consumer wants to avoid a movie with low critical acclaim, she should give priority to the opinions expressed by *Ebert*, *Mathews*, *Caro*, and *Wlosczyna*, who provide the most information in the low-acclaim range. On the other hand, a consumer with limited time or cash trying to find highly acclaimed movies should seek the opinions of *Turan*, *Ansen*, *Dargis*, *Hoberman* and *Clark* who provide the most information at the high end. Certainly, consumers may get a consensus rating that are simple averages from many different websites or newspapers, or consumers can obtain a consensus from a weighted average from www.metacritic.com. However, simple averages overlook the result that certain reviewers are more informative than others, and the Metascore's weighted average uses subjective rather than objectively measured weights. In addition, the Metascore and the simple averages

ignore what they consider to be missing information, in that a silent review will simply not be counted. Our methods provide a consensus that accounts for silence and for the degree that individual critics are informative at any particular level of acclaim.

Research on implications of "silence" in expert opinions is virtually non-existent. In any product category, a majority of products are not rated by experts or critics, simply because there are far more products than experts. As we have seen in our illustration, a majority of films are not reviewed even though the critic might have seen the movie. The prevalence of silence alone implies that it could possibly be more informative than the actual ratings of reviews themselves. For the movie data, our results confirm that the fact that an expert is silent about a product may imply a positive or a negative review, depending on the expert. Our results show that for the majority of the critics in our sample, the likelihood of a "silent" review either implies a neutral opinion or is relatively flat over the relevant range of movie acclaim (which indicates little to no opinion). However, 13 of the 46 critics studied are more likely to be silent when the movie is of low acclaim than when it is of high acclaim. This implies that silence by these reviewers is a clue that the movie might be of lower critical acclaim. Our analysis is also able to isolate a subset of 3 critics for whom silence reflects a positive opinion. We also considered the role of silence in predicting acclaim in a holdout sample of 75 movies. Our results showed that silence has a large impact on the precision of predictions, especially for highly acclaimed movies. Simply put, silence allows one to more confidently identify those movies that are of the highest acclaim.

Taken together, these results have interesting practical implications for consumers. First, if one's own preferences are in line with the first set of 13 critics, one would be better off avoiding a movie on which the critic is silent. On the other hand if one's favorite critic is in the second set of 3 critics, silence is a favorable sign, and one should consider seeing the movie on which the critic is silent. Second, and very broadly, our analyses help consumers form an overall product evaluation from diverse and conflicting opinions of experts. Third, our analysis helps consumers evaluate the experts themselves according to how informative they are at each range of the product's acclaim space. For example, our empirical illustration suggested that some movie critics are more informative than others, and some are most informative for low acclaim movies, while others provide more information in the high acclaim end.

Based on the critical acclaim measured by the model, we report a top 40 and a bottom 40, where these rankings are in the eyes of the critics alone, not by the masses that fill theatres and generate box office revenues. In a comparison of critical acclaim to box office sales, we found that acclaim is not strongly correlated with box office sales. For those who might wish to associate higher critical acclaim with higher sales, all else equal, the lack of correlation of acclaim and box office sales for the opening week here would appear troubling. The key is that "all else" is not equal. For instance, the critics and the public will likely be perfectly consistent with one another when ranking a James Bond thriller and a Schwarzenegger action film, for both were designed for mass appeal. But one should not expect mass appeal and critical acclaim to agree across wildly different movie genres, like a comparison of a James Bond thriller to a platform-released psychological study. Put another way, just as few would expect agreement between art critics and the public regarding modern art sculptures, few would expect the masses to appreciate the same movies that critics find to be of highest value (Holbrook, 1999). There are also other factors to consider in order to achieve fair comparisons across movies, to help the "all else equal" to hold more closely, such as promotion budgets, the number of screens to which the movie was released, and so on. Even so, the lack of correlation of critical acclaim and mass appeal does indicate a potential direction for future work, identifying a consensus measure for the masses that would be analogous to that we have developed for critics, ideally one that could be assessed prior to the release of the movie.

Although factor interpretability did not appear to pose a problem in our empirical illustration, a weakness of this model is that the latent factors may not have a clear interpretation, unless one can draw some intuitive interpretation from the products in the extremes of the latent space. Although our model treated positive ratings, neutral ratings, and negative ratings as nominal categories, the latent uni-dimensional acclaim measure is correlated with these ratings in the intuitive manner. For instance, the probability of a positive review increases with movie acclaim. The interpretation of latent factors can become more complex when the acclaim is multidimensional, an outcome that is feasible with our model.

Our model assumes that all movies are in the consideration set at all times, when at each period (week, or month, depending on the critic's publication cycle) the consideration set changes. Our model could be extended by defining varying consideration sets over time, which would require additional knowledge regarding the exact time of release for each movie.

References

Agresti, A., & Winner L. (1997). Evaluating agreement and disagreement among movie reviewers. *Chance, 10*(2), 10–14.

Baker, F. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann Educational Books.

Basuroy, S., Chatterjee, S., & Ravid, S.A. (2003). How critical are critical reviews? the box office effects of film critics, star-power and budgets. *Journal of Marketing, 67*, 103–117.

Bock, R.D. (1997). The nominal response model in van der Linden, Wim J. and K. H. Ronald (eds.), *Handbook of modern item response theory*, (pp. 33–49). New York: Springer.

Bradlow, E. T., & 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*(445), 43–52.

Broniarczyk, S. M., & West, P. M. (2002). Conflicting product advice: The role of prior beliefs and goals on consumer advice assessment and utilization. Working Paper, Ohio State University.

Bruce, B. (2002). Stock analysts: Experts on whose behalf? *The Journal of Psychology and Financial Markets, 3*(4), 198–210.

Eliashberg, J., & Sugan, S. M. (1997). Film critics: Influencers or predictors? *Journal of Marketing, 61*, 68–78.

Greco, A. N. (1997). *The book publishing industry*. Allyn Bacon, Needham Heights: MA.

Harvey, N., Harries, C. & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes, 81*, 252–273.

Holbrook, M. B. (1999). Popular appeal versus expert judgements of motion pictures. *Journal of Consumer Research, 26*, 144–155.

Kamakura, W. A., & Wedel, M. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika, 66*(4), 515–30.

King, M. F., & Balasubramanian, S. K. (1994). The effects of expertise, end goal, and product type on adoption of preference formation strategy. *Journal of the Academy of Marketing Science, 22*(2), 146–159.

Li, Xi, (2002). Career concerns of analysts: Compensation, termination and performance. Working Paper, Vanderbilt University.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* . Reading, MA: Addison-Wesley.

Michaely, R., & Womack, K. L. (1999). Conflict of interest and the credibility of underwriter analyst recommendations. *The Review of Financial Studies, 12*(4), 653–686.

Patz, R. J., Junker, B. W., Johnson, M. S., & 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.

Ravid, S. A. (1999). Information, blockbusters, and stars: A study of the film industry. *Journal of Business, 72*(4), 463–92.

Reddy, S. K., Swaminathan, V., & Motley, C. M. (1998). Exploring the determinants of broadway show success. *Journal of Marketing Research, 35*, 370–383.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80*(2), 159–169.

West, P. M., & Broniarczyk, S. M. (1998). Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *Journal of Consumer Research, 25*, 38–51.