YUANPING YING, FRED FEINBERG, and MICHEL WEDEL*

"Product recommendation systems" are backbones of the Internet economy, leveraging customers' prior product ratings to generate subsequent suggestions. A key feature of recommendation data is that few customers rate more than a handful of items. Extant models take missing recommendation rating data to be missing completely at random, implicitly presuming that they lack meaningful patterns or utility in improving ratings accuracy. For the widely studied EachMovie data, the authors find that missing data are strongly nonignorable. Recommendation quality is improved substantially by jointly modeling "selection" and "ratings," both whether and how an item is rated. Accounting for missing ratings and various sources of heterogeneity offers a rich portrait of which items are rated well, which are rated at all, and how these processes are intertwined, while reducing holdout error by 10% or more. The authors discuss ways to implement the proposed framework within existing recommendation systems and its implications for marketers.

Leveraging Missing Ratings to Improve Online Recommendation Systems

Aiming to be "the Earth's most customer-centric company," Amazon.com is engaged in an ongoing effort to refine its online personalization capabilities. Chief among these is the ability to offer product recommendations based on customers' prior behavior on the Web site. Using product recommendation software developed by NetPerceptions, Amazon.com compares the target customer's browsing and purchasing profile with those of other customers and uses other customers' product evaluations to suggest what the target customer might like. This personalization capability of turning customer knowledge into product recommendations and, ultimately, into purchases is key to the company's customer retention strategy.

Amazon.com is not alone in this endeavor. CDNow uses a similar system to recommend recordings, Blockbuster makes personalized video recommendations to its customers, Netscape incorporates its "What's Related?" capability to evaluate commonalities in Web site visiting behavior, Netflix asks its customers to e-mail ratings for movies they have viewed as a basis for its recommendations, and so on. An industry has emerged to offer turnkey solutions to firms hoping to match customers with the "right" product by leveraging the histories and product evaluations of prior customers.

This article examines the development of "product recommendation systems." For several reasons, making personalized product recommendations has become an important goal for marketers, not only those operating through the newer electronic channels but also those operating through traditional bricks-and-mortar channels. First, by recognizing customer heterogeneity and capitalizing on similarities in preference, product recommendation systems enable companies to cater to individual preferences. Second, they allow marketers to recommend products to a target customer using information from other customers with similar product preferences, prior purchase histories, and/or demographic profiles. Recommendation systems enable marketers to act as word-of-mouth agents and virtually expand customers' social networks. Third, accurate product recommendations save customers search costs during the purchase decision process. This is important if customers face a large number of choice alternatives, particularly if the product's attributes are difficult to evaluate before consumption or are not helpful in anticipating the product's consumption utility. Examples include hedonic products, such as music, movies, books, restaurants, tourism, and other products for which there are substantial experiential components. For making a choice among such hedonic products, other customers' evaluations are particularly

*Yuanping Ying is Assistant Professor of Marketing, School of Management, University of Texas, Dallas (e-mail: yingyp@utdallas.edu). Fred Feinberg is Hallman Fellow and Bank One Corporation Associate Professor of Marketing, Stephen M. Ross School of Business, University of Michigan (e-mail: feinf@umich.edu). Michel Wedel is PepsiCo Professor of Marketing, Robert H. Smith School of Business, University of Maryland (e-mail: mwedel@rhsmith.umd.edu). The authors thank Peter Lenk for his helpful suggestions.

© 2006, American Marketing Association
ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. XLIII (August 2006), 355–365

important, and customers actively welcome product recommendations based on others' experiences. Finally, an effective product recommendation system can help with customer acquisition and retention and enhance customer loyalty. Thus, it has become an important tool in customer relationship management.

In this article, we develop a statistical framework for improving product recommendations, building on previous work by Ansari, Essegaier, and Kohli (2000). We view product recommendations as predictions of a target customer's latent consumption utility; that is, products for which the customer has a high predicted utility should be recommended. In doing so, we address three problems that have been neither fully nor jointly resolved by current algorithms: missing data, the ordinal nature of the rating scales and their use, and customer heterogeneity.

In our view, the first of these problems, which pertains to how products come not to be rated, is the most serious, and it is the main focus of this article. Most of the product-rating data on which recommendation systems rely are not only missing but also missing nonrandomly. Customers can evaluate only products they have experienced, so they commonly rate only a small subset of all available items. Consequently, compared with the entire product catalog an e-tailer offers, the ratings history of any particular customer is woefully sparse, even within a single product category. A customer is rarely familiar with even a small fraction of Netflix's films, let alone Amazon.com's book catalog. Therefore, some current recommendation systems overtly request that customers evaluate products or provide preferences, and indeed, such information can be broadly effective in generating subsequent recommendations. For example, Art Technology Group's Dynamo server requests customers' preference input and has established clientele, such as Blockbuster, J.Crew, and Target. Customers are usually asked to supply evaluations for only a few products from among those they are familiar with because evaluation of the complete set of alternatives is neither feasible nor desirable. This can only amplify the sparseness of the product recommendations data and their selective nature.

Therefore, the chief task for recommendation systems is to predict a target customer's evaluations using available ratings data from that customer and numerous others. In doing so, possible causes of the missing-data pattern should be considered. Previous product recommendation systems have been based only on observed ratings, tacitly presuming that missing data are valueless and thus can be safely ignored. Although the assumption that missing data lack useful information content seems reasonable on the surface, the assumption is valid only if the missing evaluations are missing completely at random (see Rubin 1976); in other words, the fact that an evaluation is missing should not depend in any way on the value of that evaluation. This would be a reasonable basis for model building if customers failed to provide product preferences for reasons not systematically related to variables relevant to the study, but this is unduly restrictive and runs counter to intuition. For example, some customers do not offer ratings simply because they have no consumption experience with the products in question, and in turn, the very reason that customers do not purchase or consume the products in question may be that they simply do not like them. Even if customers have consumption experience with certain products, they are not necessarily willing to provide marketers with their evaluations. Some customers may offer their views only of products that they either like or dislike very much to champion some while giving warning about others. Web sites such as Epinions.com have sprung up to facilitate exactly this sort of product evaluation sharing.

In summary, aside from parsimony, there is little reason to believe that missing evaluations are missing completely at random, lacking a useful pattern from which statistical models can extract meaning. Because it is well known that failing to accommodate nonignorable missing-data mechanisms can lead to biased estimates (Little and Rubin 1987), it is possible that this previously unspoken assumption could lead to suboptimal or even erroneous recommendations. Therefore, one goal of this article is to model the missing-data mechanism explicitly, thus alleviating the problem of biased estimates and improving recommendation quality. Although we cannot explore the specific causes underlying the missing-data pattern, as we discuss subsequently, it is clear that the data should not be ignored.

The second problem pertains to the ordinal nature of most product evaluation data that are input to recommendation algorithms. Treating the evaluation data as either nominal or interval scaled, which several prior models have done, fails to reflect the true data generation mechanism and thus potentially offers inconsistent forecasts. We specifically model ratings data as ordinal, subject to estimated cutoffs, and we account for this ordinality in our model comparisons. Permitting these estimated cutoffs to vary across customers simultaneously allows for some degree of scale usage heterogeneity.

The third problem is the heterogeneity that is inherent in customers' preference data. Although marketing researchers have used both finite mixture and hierarchical Bayes models to capture customer heterogeneity, several studies suggest that these models offer roughly equal performance in various empirical settings (Andrews, Ansari, and Currim 2002; Wedel et al. 1999). For the purpose of recommendation systems, hierarchical Bayes models offer a compelling method for modeling individual-level heterogeneity, which is particularly useful in the context of one-to-one marketing (Allenby and Rossi 1999). Therefore, we adopt the hierarchical Bayes approach, which allows for individual-level posterior parameter estimates. As we discuss subsequently, these individual-level posteriors offer superior recommendation quality for all the model types we explore.

In the next section, we provide a review of prior research and approaches to making individualized recommendations. Then, we present the general statistical framework for product recommendations and delineate our model. The following section describes a range of empirical tests, in which we reinvestigate the benchmarking EachMovie data set that has been used in a great deal of prior research on recommendation systems. We compare our approach with several popular alternatives and restricted variants, and we conclude with a discussion of future research directions.

RECOMMENDATION SYSTEMS LITERATURE REVIEW

Although recommendation systems have been a popular topic of study in the computer science and machine-learning literature, only recently has the marketing litera-

ture made this a core topic of study, owing to the pioneering work of Ansari, Essegaier, and Kohli (2000). We focus on recommendation algorithms of a statistical nature; for a broader history of recommendation systems, see Ansari, Essegaier, and Kohli's comprehensive overview.

The central research problem encountered in making product recommendations can be construed as predicting the entries of Table 1, which represent customers' ratings for a given set of products, along with missing data for the products not rated. As we argue throughout the article, both types of data should be modeled simultaneously. Product ratings are typically made on preordained ordinal scales. For example, Amazon.com, Netflix, and Blockbuster each ask customers to award products between one and five "stars," and the EachMovie data we analyze have a similar star rating.

Existing recommendation methods can be categorized into two classes: heuristic methods and model-based methods. Heuristic methods (often clustering-type algorithms, including nearest-neighbor methods) are widely used in the computer science literature, in which researchers have attempted to filter out irrelevant information from what is available on the Internet. The popularity of these heuristic approaches stems from their ease of implementation, but they are often ad hoc and have been shown to be broadly inferior to model-based methods (Breese, Heckerman, and Kadie 1998). Model-based methods invoke a probability distribution for customers' responses and therefore explicitly hypothesize a data generation process. Model-based methods that have been used to generate product recommendations include the mixture model (Chien and George 1999), the hierarchical Bayes model (Ansari, Essegaier, and Kohli 2000), factor analysis (Canny 2002), and Bayesian network models (Breese, Heckerman, and Kadie 1998).

Chien and George (1999) were the first to propose a powerful recommendation system based on a Bayesian mixture model. Their approach can be readily implemented, and they show that it outperforms nearest-neighbor methods (Sarwar et al. 2000) on the EachMovie data. Because much subsequent research has taken its cue from Chien and George's pioneering work, it is instructive to discuss how it might be extended. First, although the EachMovie customer preference data are ordinal, Chien and George treat them as nominal. Although this allows for a great deal of flexibility (e.g., in accommodating different "patterns" of response across the nominal categories), it cannot capitalize on the ordinal relationships that are intrinsic to the true data generation process, and as a practical concern, it is not very parsimonious. Second, Chien and George do not include explanatory variables to help elucidate why customers may like or dislike a product. Consequently, their model cannot predict customers' preferences for new products, for which detailed attribute-level data are the only information available. Indeed, we view predictions for new products among the chief overall benefits of recommender systems, and we discuss them at length subsequently. Third, the mixture approach assumes that all customers in a particular segment have the same preference structure. Although this is intuitively appealing, it can be restrictive in practice and entail a large number of parameters, requiring substantial computation time and potentially reducing holdout predictive validity. Finally, as in other current recommendation systems, and most important from the perspective of the current study, Chien and George assume that data are missing completely at random, and thus they do not posit a mechanism for the (nonignorable) missing data.

Ansari, Essegaier, and Kohli (2000) propose an effective hierarchical Bayes model to predict customers' ratings for products, allowing for both fixed effects and random effects of customer characteristics and of product attributes. Specifically, let X_i denote customer i's characteristics, and let Z_j denote product j's attributes. Then, customer i's rating for product j, Y_{ij}, is modeled as follows:

$$(1)\qquad \begin{aligned} Y_{ij} &= \beta^X X_i + \beta^Z Z_j + CH_i + PH_j + e_{ij}, \\ CH_i &= \lambda_{i1} + \lambda_{i2} Z_j, \text{ and} \\ PH_j &= \gamma_{j1} + \gamma_{j2} X_i, \end{aligned}$$

where $e_{ij} \sim N(0, \sigma^2)$, $\lambda \sim N(0, \Lambda)$, and $\gamma \sim N(0, \Gamma)$. Heterogeneity is captured by two model components: CH_i for customers and PH_j for products. Their model is also tested on the EachMovie data. Ansari, Essegaier, and Kohli's (2000) approach is elegant and practicable, though there are a few caveats. Rather than mimicking the ordinal nature of customer preference data, they treat customers' product ratings as if they are measured on interval scales. They implement a post hoc grid search to translate these continuous latent ratings into observable discrete ordinal ones, though it is possible to estimate these simultaneously with other model parameters. Similar to Chien and George (2000), Ansari, Essegaier, and Kohli do not consider the vast proportion of missing preference ratings.

In summary, although both the mixture and the hierarchical Bayes approaches have proved to be powerful model-

Table 1
SAMPLE CUSTOMER PREFERENCE DATA

	Products						
Customers	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	…
C1	5	Missing	1	3	Missing	4	…
C2	Missing	Missing	6	2	3	Missing	…
C3	4	2	Missing	Missing	Missing	6	…
C4	1	5	3	Missing	6	3	…
C5	6	Missing	6	Missing	2	Missing	…
C6	Missing	Missing	2	5	4	Missing	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

based methods to generate product recommendations, neither has jointly accommodated all three key features of customer preference data that we focus on in this study: nonignorable missing data, the ordinal nature of ratings scales, and a fairly general account of heterogeneity. In the next section, we present a unified approach that captures all these features, and we verify that it can indeed enhance the quality of product recommendations.

THE RECOMMENDATION MODEL

In line with prior research, we propose a product recommendation algorithm based on ratings-based customer preference data. We assume that customers' ratings for products across different product categories are elicited using an ordinal scale with K categories.

We denote the observed rating as Y_{ij}, where $Y_{ij} = k$ if customer i rates product j as k (k = 1, …, K). Customers differ in the number of products they rate, which results in an unbalanced data set. The model allows for two covariate sets: information on customer characteristics and on product attributes. Our objective is to predict (see Ansari, Essegaier, and Kohli 2000) (1) how an existing customer would have rated an existing product that he or she has not rated, (2) an existing customer's rating for a new product, (3) a new customer's rating for an existing product, (4) a new customer's rating for a new product, and (5) probabilities of rating an item at all for each of these customer × product groups. For the first four points, we compare ratings that stem from what we term the "prediction model" component with customers' actual ratings; that is, these comparisons are relative to the set of products that were actually rated. For the fifth point, our approach affords the unique advantage of using the "selection model component" to assess the much larger set of items that were not rated (i.e., not selected for rating). Throughout, we maintain a strict separation between predicting a "rating," which refers to registering an evaluation for an item on a fixed ordinal scale, and "selection," which refers to whether an evaluation is registered at all. Note that we account for these on two separate scales: an ordinal scale for rating and a binary scale for selection; as such, we take care to use appropriate and distinct measures to assess their accuracy.

Joint Model for Ordinal Ratings and Binary Selection

Missing product ratings are ubiquitous in data collected for recommendation systems, and previous research on the topic implicitly considers them missing completely at random. In actuality, as we noted previously, customers may fail to provide ratings for certain products for numerous reasons, such as lack of awareness or consumption experience, tendency to rate only strongly liked or disliked products, or the recommendation system itself presenting a preselected set of products to customers for rating. Thus, the missing-completely-at-random assumption is unlikely to provide a faithful description of the process by which the missing data are generated, and thus we relax it by explicitly incorporating a missing-data mechanism in our model. We emphasize up front that it is hazardous to interpret the model we present subsequently as one that reflects a two-stage process of customer behavior. The reason for this note of caution is that we are necessarily unaware of the range of potential causes underlying the selection process. That is, it may be the case (e.g., based on prior ratings) that the recommendation system itself picks products for customers to rate, that customers select which products they rate of their own volition, or a mixture of both. The selection component of the model can account for such a variety of causes for missing ratings, but in absence of specific data or knowledge of exactly how selection comes about, it does not allow us to distinguish a selection mechanism driven by the system itself from one driven solely by the behavior of its users.

The key idea here is that knowledge that a product was rated (hereinafter, "selected") should influence the analyst's estimate of the value of that rating (hereinafter, "prediction"). To account for this, we propose a joint model for the latent processes underlying selection and prediction. Subsequently, we provide the model in full detail. However, we begin by outlining the general model structure, suppressing subscripts for products and customers and using the generic symbol X to subsume all covariates. This helps clarify how we model the ratings and selection components and the relationships between them, and it helps show how recommendations follow from the model. After describing the data, we provide specifics of the model, incorporating heterogeneity in all parameters and distinguishing between customer characteristics and product attribute covariates in X. The model structure is given by

$$(2)\qquad U_s = \beta_s X_s + \varepsilon_s$$

$$U_p = \beta_p X_p + \varepsilon_p$$

$$(\varepsilon_s, \varepsilon_p) \sim N(0, 0, 1, 1, \rho).$$

The latent model can be compactly described as $(U_s, U_p) \sim N(\beta_s X_s, \beta_p X_p, 1, 1, \rho)$; note that model identification dictates the unit variances of the error specification. We must specify mechanisms for translating these latent values into observables and opt for an especially flexible method, using estimated (heterogeneous) cutoffs. Although some formulations for ordinal data are more parsimonious, such parsimony typically comes at the cost of imposing some shape on the ratings distribution. For example, Rost's (1985) specification allows an entire ordinal distribution to be compressed into a single parameter, which determines its mean, variation, and tail properties and imposes unimodality. Because our focus is on enhancing recommendation quality, we retain flexibility in accounting for the distribution of ratings across observed ordinal categories 1, …, K. Therefore, we assume that latent values are translated into observed quantities as follows:

$$(3)\qquad \Pr(Y_s = 1) = \Pr(0 < U_s)$$

$$\Pr(Y_p = k) = \Pr(\kappa_{k-1} < U_p \le \kappa_k).$$

The system in Equations 2 and 3 comprises a binary probit mechanism for whether a rating is observed (selection) and an ordinal probit mechanism for which rating is observed (prediction). The correlation, ρ, which we allow to vary across individuals, captures the interrelation between the selection and the prediction portions of the model. The cutoffs, $\{\kappa_0, \ldots, \kappa_K\}$, determine how the latent scale is mapped onto the K observed ordinal scale points. For iden-

tification purposes, we set the two lowest cutoffs $\{\kappa_0, \kappa_1\}$ to negative infinity and zero, respectively; we set κ_K to infinity; and we estimate the remaining cutoffs $\{\kappa_2, \ldots, \kappa_{K-1}\}$. Thus, for K ordinal response categories, we estimate K – 2 cutoffs. Although for parsimony these are often taken as constant across customers and products, here we allow them to be heterogeneous, along with $\{\beta_s, \beta_p, \rho\}$, as we describe in the following section.

Making Recommendations for Specific Products

Although no assumptions need to be made about temporal ordering, the formulation of Equations 2 and 3 might be interpreted as being consistent with a two-stage ratings provision process: First, a decision is made (based on the rating system's request, customer initiative, or both) about whether to provide a product rating, according to the selection model; second, if the decision is affirmative, the rating is recorded. Specific product recommendations stem from the model as follows: We need to know the expected value for the prediction utility, $\beta_p X_p + \varepsilon_p$. When the stochastic component is unrelated to any other modeled quantities, this is merely $\beta_p X_p$. However, when we know that a rating is observed, we must condition on this fact, so the expected observed rating is

$$
\begin{aligned}
(4)\qquad & E(\beta_p X_p + \varepsilon_p | \beta_s X_s + \varepsilon_s > 0) \\
&= \beta_p X_p + E(\varepsilon_p | \varepsilon_s > -\beta_s X_s) \\
&= \beta_p X_p + \frac{1}{\Phi(\beta_s X_s)} \int_{u=-\beta_s X_s}^{u=\infty} \rho u \phi(u) du \\
&= \beta_p X_p + \rho \frac{\phi(\beta_s X_s)}{\Phi(\beta_s X_s)}.
\end{aligned}
$$

We use this quantity, involving the inverse Mills ratio, $\varphi(\cdot)/\Phi(\cdot)$, in predictions that involve all comparisons of observed recommendations for our model.

Heterogeneity and Explanatory Variables

It is crucial to account for heterogeneity across customers. Because we wish to enhance the quality of these one-to-one recommendations, we opt for the hierarchical normal heterogeneity specification, as Ansari, Essegaier, and Kohli (2000) use. In this way, we account for heterogeneity in the effects of predictor variables on the ratings and for heterogeneity in the usage of the rating scales. In addition, we not only model heterogeneity in the ratings data in this manner but also account for heterogeneity in the selection process generating the missing product data. Thus, we obtain posterior estimates of the selection coefficients, prediction coefficients, error correlation, and cutoffs, $\{\beta_s, \beta_p, \rho, \kappa\}$, and use them to obtain individual-specific recommendations through Equation 4.

Because ρ is bounded in [–1, 1], we allow for normal heterogeneity in the inverse hyperbolic arctangent of ρ instead; note that $\text{atanh}(\rho) = (1/2)\ln([1 + \rho]/[1 - \rho])$ maps [–1, 1] to the real line and is a common transform for correlation. Similarly, we cannot impose normal heterogeneity on the cutoffs directly, because they are positive and obey order restrictions. Rather, we appeal to a transformation to the whole real line and take logs of adjacent cutoff differences, $\Delta\kappa = \{\kappa_1 - \kappa_0, \ldots, \kappa_K - \kappa_{K-1}\}$, to be distributed normally across customers, with a full covariance matrix, as specified precisely for all parameters in the following section.

Full Model Specification

With i denoting customers, j denoting movies, s denoting selection, p denoting prediction, G denoting genres (for movies, j), and D denoting demographics (for Customers, i), the model is given in full as follows:

$$
\begin{aligned}
(5)\qquad & U_{ijs} = \beta^G_{is} G_{js} + \beta^D_{js} D_{is} + \varepsilon_{ijs} \\
& U_{ijp} = \beta^G_{ip} G_{jp} + \beta^D_{jp} D_{ip} + \varepsilon_{ijp} \\
& \beta^G_{is} \sim \text{MVN}(\mu^G_s, \Delta^G_s),\ \beta^G_{ip} \sim \text{MVN}(\mu^G_p, \Delta^G_p) \\
& \beta^D_{js} \sim \text{MVN}(\mu^D_s, \Delta^D_s),\ \beta^D_{jp} \sim \text{MVN}(\mu^D_p, \Delta^D_p) \\
& \log(\Delta\kappa_i) \sim \text{MVN}(\mu_c, \Delta_c) \\
& \tanh^{-1}(\rho_i) \sim N(\mu_\rho, \sigma^2_\rho) \\
& (\varepsilon_{ijs}, \varepsilon_{ijp}) \sim \text{BVN}(0, 0, 1, 1, \rho_i).
\end{aligned}
$$

Note that coefficients for genres are heterogeneous across customers i, whereas those for demographics must be taken as homogeneous across customers because they do not vary across observations for a particular customer; however, they are heterogeneous across movies, j. All estimated covariance matrices $\{\Delta^G_s, \Delta^G_p, \Delta_c\}$ are full (i.e., nondiagonal). Because each of the selection and prediction models can have only one estimated intercept mean, the first elements in μ^D_s and μ^D_p are set to zero.

EMPIRICAL ANALYSIS

The EachMovie Data

Compaq Equipment Corporation provided the data for our analysis. Data were collected through a movie recommendation system called EachMovie. Because much previous research has used the same data source (see, e.g., Ansari, Essegaier, and Kohli 2000; Breese, Heckerman, and Kadie 1998; Chien and George 1999; Hofmann and Puzicha 1999), the EachMovie data provide an excellent benchmark to gauge relative model performance.

The data set includes 72,916 customers' numerical ratings for 1628 different movies, collected between March 1996 and September 1997. We coded the ratings on a six-star scale (K = 6). As might be expected, any particular customer was unlikely to rate even a moderate proportion of all available movies, resulting in an enormous number of missing values in the 72,916 × 1628 matrix implied by Table 1. The data set also includes information on customer demographics and the movie genre. In the subsequent analysis, genre (G_j) appears as a product-level nominal descriptor variable (action, animation, art/foreign, classic, comedy, drama, family, horror, romance, and thriller), and demographics (D_i) consist of age and gender for each customer.

For model calibration, we draw a random sample of 2432 customers from among those who provided complete demographic information. To ensure stability in model estimation, we take a random sample of 78 movies (from among

those rated at least once). Still, our calibration sample is extremely sparse, with 93.9% missing values (we set up our other samples to keep this "missingness" rate close to that of the calibration sample). Analogous to Ansari, Essegaier, and Kohli's (2000) work, we construct three holdout samples for various prediction purposes, obtained as random samples of the eligible customers and movies; we detail these samples subsequently.

Estimation

We estimate the model using Markov chain Monte Carlo methods. Given the data and dimensions described previously, we set priors to

$$\{\mu_s^G, \mu_s^D, \mu_p^G, \mu_p^D\} \sim \text{MVN}(0, 100I) \tag{6}$$

$$\mu_c \sim \text{MVN}(\mathbf{0}, 10I),\ \mu_\rho \sim \text{N}(0, 10)$$

$$(\Delta_s^G)^{-1},\ (\Delta_p^G)^{-1} \sim \text{Wishart}(0.1I, 15);\ \Delta_s^G,\ \Delta_p^G \text{ each } 11 \times 11$$

$$(\Delta_s^D)^{-1},\ (\Delta_p^D)^{-1} \sim \text{Wishart}(0.1I, 6);\ \Delta_s^D,\ \Delta_p^D \text{ each } 3 \times \mathbf{3}$$

$$(\Delta_c)^{-1} \sim \text{Wishart}(0.1I, 6);\ \Delta_c \text{ is } 4 \times 4$$

$$(\sigma_\rho^2)^{-1} \sim \text{Gamma}(0.001, 0.001).$$

For all model results we report in this article, we discard the first 20,000 draws for burn-in and use at least 30,000 additional draws to characterize the posterior distributions of the parameters. Analysis of several synthetic data sets supports the performance of the algorithm in recovering "true" parameters and reveals that convergence is reached well before the end of the burn-in. We assess convergence by inspecting plots of draws against iterations, as well as several convergence statistics, such as the Gelman-Rubin diagnostic (Brooks and Gelman 1998). Model results are summarized through the posterior means and standard deviations of the parameters.

Model and Results

We estimate the model of Equations 2 and 3 on the calibration data set; estimates appear in Table 2. The richness of the hierarchical specification allows us to explore several types of differences systematically (1) across genres, (2) by demographic groups, (3) across individuals, and (4) between selection and rating. Furthermore, we compare several models for three types of holdout samples introduced subsequently.

Differences across genres. There are strong and intuitive differences across genres in terms of how well each is liked (rated) overall. If we hold aside differences across customers, it is clear that neither family-oriented movies nor thrillers are well liked by the panelists. In contrast, animation, art/foreign, and classic movies are well received on average. If a model were formulated for ratings alone, these might well be the final conclusions. However, note that though classic movies are rated highly, they are selected less frequently than any other genre type, by a wide margin. This may indicate a core constituency of classic movie devotees who, though small in number, prize classic movies very highly, or it may reflect that EachMovie tends not to ask for ratings of these older movies. We elaborate on differences between selection and ratings subsequently.

Differences by demographic groups. Because the model specification includes demographics (age: standardized; gender: contrast coded), we can explore how each affects selection and ratings. Demographic patterns are broadly concordant for selection and ratings. For example, men are slightly less likely to select a movie but are more likely to rate it highly than women. Age yielded a per-year effect of –.046 for selection and .035 for ratings (though the latter falls short of significance). Because we standardized error variances for the latent variable model to one, each of these per-year figures is fairly large. Apparently, older customers are progressively less likely both to select a movie (for rating) and to rate it highly.

Differences across individual customers. Because the normal account of heterogeneity summarizes cross-customer differences in a single quantity, $(\text{diag}[\Delta])^{1/2}$, it is simple to spot variables across which these differences are relatively large. In terms of selection behavior, we observe the greatest variation across some of the least popular (i.e., least selected) genres, including classic, thriller, and family. We observe a rather different pattern in terms of ratings, with the greatest variation across horror, art/foreign, drama, and classic (which were not among the least well rated). Certain genres, such as comedy and animation, seemed to vary little in terms of either selection or rating behavior, indicating that they may be perceived similarly across the customer base.

Differences between selection and rating. Even when demographic differences are not considered, differences between whether a film is rated and how it is rated are stark. This is perhaps clearest for both classic and art/foreign films; neither is apparently very popular (considering how often they are selected), and both have negative coefficients (though there is a good deal of variation, per the relative magnitude of their random effects). However, despite how infrequently they are chosen, the ratings achieved by these movies in these two genres are the highest overall (except for the ever-popular animation genre). Simply put, two of the genres selected least are rated the best. Although this sort of pattern is common for luxury goods in many categories, we could not have anticipated it for movies, because classics are widely available (unlike art/foreign films) and are often less costly on a per-unit rental basis. In contrast, animated films are both highly selected and highly rated, each with low variation across customers, whereas comedy, drama, and horror fall near the middle of the pack on both. The evident pattern of differences between selection and rating behavior is not easily captured with a single epithet and may partly reflect EachMovie's policy to solicit ratings from its customers. When rating and selection behavior are broadly concordant, they do not need to be modeled separately, so that an account of ratings alone (as offered by prior models) would suffice. However, such a pattern is not apparent for the EachMovie data.

Holdout Recommendation Validity

The results of the previous section help elucidate what the proposed model reveals about drivers of both selection and ratings. Although this is of interest in itself, the acid test of any recommendation system is how it makes actual recommendations. Thus, we must examine recommendation

Table 2
ESTIMATES OF GENRE AND DEMOGRAPHIC EFFECTS ON SELECTION AND RATINGS

	Mean Effects		*Random Effects*	
Selection Coefficients	μ	*SE*	$(Diag[\Delta])^{1/2}$	*SE*
Heterogeneous Across Customers				
Intercept	–2.802	.091	.532	.028
Genre 1 (action)	.854	.122	.346	.027
Genre 2 (animation)	.809	.109	.345	.072
Genre 3 (art/foreign)	–.523	.101	.362	.050
Genre 4 (classic)	–2.070	.128	1.118	.093
Genre 5 (comedy)	.233	.067	.263	.027
Genre 6 (drama)	.117	.120	.279	.025
Genre 7 (family)	–.585	.119	.888	.049
Genre 8 (horror)	.395	.130	.398	.042
Genre 9 (romance)	.594	.115	.232	.034
Genre 10 (thriller)	–.607	.117	.602	.048
Heterogeneous Across Movies				
Intercept	0	0	.912	.077
Age[a]	–.046	.024	.122	.016
Gender[b]	–.022	.050	.170	.024
Rating Coefficients				
Heterogeneous Across Customers				
Intercept	.249	.149	.762	.057
Genre 1 (action)	.772	.095	.445	.047
Genre 2 (animation)	2.110	.111	.389	.090
Genre 3 (art/foreign)	1.462	.185	.809	.093
Genre 4 (classic)	1.375	.094	.666	.190
Genre 5 (comedy)	.531	.080	.415	.041
Genre 6 (drama)	.506	.057	.698	.050
Genre 7 (family)	–.389	.142	.263	.044
Genre 8 (horror)	.312	.160	.905	.057
Genre 9 (romance)	.562	.118	.510	.062
Genre 10 (thriller)	–.053	.213	.188	.057
Heterogeneous Across Movies				
Intercept	0	0	.825	.092
Age[a]	.035	.037	.176	.027
Gender[b]	.123	.059	.176	.039
Correlation and Cutoffs (as Estimated)[c]				
atanh(ρ)	.106	.042	.235	.036
log($\kappa[2]$)	–1.131	.055	.411	.076
log($\kappa[3] - \kappa[2]$)	–.578	.029	.264	.031
log($\kappa[4] - \kappa[3]$)	–.127	.020	.253	.030
log($\kappa[5] - \kappa[4]$)	.122	.020	.260	.030
Correlation and Cutoffs (Transformed from Above)				
ρ	.105	.042		
$\kappa[2]$	.323	.018		
$\kappa[3]$	.884	.022		
$\kappa[4]$	1.766	.027		
$\kappa[5]$	2.896	.038		

[a]Age is standardized.
[b]Female is coded as –½, and male is coded as ½.
[c]$\kappa[0] = -\infty$, $\kappa[1] = 0$, and $\kappa[K] = \infty$.

performance in holdout samples and comparisons with other models. To that end, we compare model performance with Ansari, Essegaier, and Kohli's (2000) hierarchical Bayes model as a benchmark.[1] A modified version of Ansari, Essegaier, and Kohli's original model must be estimated because the experts' ratings they used as explanatory variables are not available for all movies in our data. Thus, for comparison purposes, explanatory variables are the

[1]See Table 3 in their article. We also estimated and compared results for Chien and George's (1999) model. Because it makes categorical, not ordinal, predictions and because it can be used for only one of our holdout samples (when it is dominated by the other two methods), we do not include these results, but they are available on request.

same as those we have used thus far: age and gender for customers and the ten genres for movies.

It is convenient to view the calibration sample as "existing customers/existing movies," so that there are three possible prediction or holdout samples: "new customers/new movies," "existing customers/new movies," and "new customers/existing movies." The latter two are particularly telling because they allow us to gauge the value of leveraging the vast base of movie ratings data to a new set of customers and the relatively shallow, though individual-specific, base of ratings data to a new set of movies, respectively. The new customers/new movies holdout set offers a stringent test because there would be no a priori reason to expect any model to do especially well, relying only on genre and demographic information to describe the ratings pattern of a new set of customers for a new set of movies. Table 3 provides descriptive statistics of the calibration sample and the three holdout samples. We obtained each of the holdout samples through random sampling of eligible customers and movies. To provide the fairest test against Ansari, Essegaier, and Kohli's (2000) model, which did not specify a missing-data mechanism, we strove to keep the proportion of missing data in the holdout samples approximately the same as in the calibration sample.

We report the holdout recommendations for our model and Ansari, Essegaier, and Kohli's (2000) model in Table 4. The results are segregated into two types: movies that were selected and those that were not. It is important to realize that these are not comparable; an unrated movie did not receive a "rating of zero." Whether a movie was rated is purely binary and reflects the performance of the selection model component. What rating a movie received, conditional on having been selected, lies on an ordinal scale and reflects the performance of the prediction model component. Therefore, we present five different measures, two that are appropriate for binary (selection) data and three for

Table 3
SUMMARY OF SAMPLES

			Age			
Samples	*Number of Customers*	*Number of Movies*	*M*	*SD*	*Female (%)*	*Missing (%)*
Existing customers/existing movies[a]	2432	78	31.8	12.0	21.8	93.9
Existing customers/new movies	1913	73	31.4	11.7	21.7	93.2
New customers/existing movies	2339	76	32.0	11.8	20.8	93.6
New customers/new movies	2009	78	31.9	11.7	21.1	93.9

[a]Calibration sample.

Table 4
HOLDOUT RECOMMENDATION RESULTS

Prediction: Ratings Accuracy for Selected Movies									
	New Customers/ New Movies			*Existing Customers/ New Movies*			*New Customers/ Existing Movies*		
Models	*MAD*	*RMSE*	*Spearman*	*MAD*	*RMSE*	*Spearman*	*MAD*	*RMSE*	*Spearman*
Joint model, heterogeneous ($\rho \neq 0$)	**1.190**	**1.529**	.147	**1.178**	**1.561**	.140	**1.037**	**1.436**	**.346**
Joint model, homogeneous ($\rho \neq 0$)	1.262	1.638	.123	1.249	1.621	.147	1.131	1.512	.220
Prediction model, heterogeneous ($\rho = 0$)	1.236	1.560	**.155**	1.219	1.581	**.208**	1.062	1.470	.339
Prediction model, homogeneous ($\rho = 0$)	1.262	1.638	.123	1.249	1.621	.147	1.131	1.512	.220
Ansari, Essegaier, Kohli (2000)	1.440	1.760	.075	1.347	1.685	.037	1.069	1.403	.316

Selection: Mean Probabilities for Selected Movies			
Joint model, heterogeneous ($\rho \neq 0$)	6.64%	**9.50%**	**22.71%**
Joint model, homogeneous ($\rho \neq 0$)	**9.13%**	9.02%	14.03%
Selection model, heterogeneous ($\rho = 0$)	6.46%	8.67%	19.49%
Selection model, homogeneous ($\rho = 0$)	7.20%	7.20%	12.30%
Sample nonmissing rate	6.10%	6.85%	6.43%

Hit Rate: (Number of Correctly Predicted Ratings/Total Number of Observations)			
Joint model, heterogeneous ($\rho \neq 0$)	**25.95%**	**27.75%**	**32.34%**
Joint model, homogeneous ($\rho \neq 0$)	25.38%	25.72%	29.13%
Selection model, heterogeneous ($\rho = 0$)	24.16%	25.65%	31.96%
Selection model, homogeneous ($\rho = 0$)	25.38%	25.72%	29.13%
Ansari, Essegaier, Kohli (2000)	18.89%	22.30%	28.50%

Notes: Optimum value across all methods in each column appears in bold.

ordinal (prediction) data; specifically, we assess mean probabilities and hit rates for binary selection and mean absolute deviation (MAD), root mean square error (RMSE), and Spearman rank-order correlation for ratings predictions.[2] Recall that Ansari, Essegaier, and Kohli's model yields continuous predictions; using integer programming, we replicate the grid search procedure outlined in their work to select optimal grid cutoff points to translate the continuous predictions into discrete ratings. Moreover, because their model does not account for missing data, we compare selection probabilities for our model alone with various restricted variants.

The summary statistics of Table 4 enable us to compare and contrast the estimated models in several ways, specifically, in terms of the marginal explanatory power of (1) movie genre and customer demographic covariates, (2) customer and movie heterogeneity, (3) correlation between selection and prediction, and (4) the proposed model compared with restricted variants and Ansari, Essegaier, and Kohli's (2000) benchmark. We consider these in turn, comparing performance in terms of both prediction and selection across the three different samples (i.e., new customers/new movies, existing customers/new movies, and new customers/existing movies). Increasing performance might be expected across these samples because there are far more data for each of the movies than for any individual customer.

Covariates for movie genre and customer demographics. That covariates are useful in understanding both prediction and selection is apparent from the detailed results of Table 2; with two minor exceptions, all have strongly significant means, and most have relatively tight random effects distribution. We also compare a "pure," no-covariates selection model, as embodied by the "sample nonmissing rate," with the other four possibilities: the selection-only ($\rho = 0$) and the joint models, each with and without heterogeneity. Taken as a set, the genre and demographic covariates boost the mean probability of the selected movies in all cases, strongly so for the new-customers/existing-movies sample (from a baseline of 6.4% to 12.3% for the homogeneous selection model and to 14.0% for the homogeneous joint model). Heterogeneity (in movies, for this sample) offers dramatic further improvements to 19.5% and 22.7%, respectively. A similar pattern is evident for hit rates, which increase from a baseline of 28.5% to 32.3% for the joint heterogeneous model.

Customer heterogeneity. The importance of accounting for customer heterogeneity is apparent in the existing-customers/new-movies sample, in which customer-specific posterior values of coefficients, correlations, and cutoffs can be called on for both selection and prediction. With one exception, customer heterogeneity "grafted on" to any model improves its associated fit measures. In terms of the joint model, customer heterogeneity reduces (improves) MAD for the joint model from 1.249 to 1.178, increases selection probability from 9.0% to 9.5%, and boosts hit rate from 25.7% to 27.8%, all greater than other model effects.

Correlation between selection and prediction. The key construct in the proposed joint model is the correlation between selection and prediction. Without such a correlation, selection and prediction are independent, and neither can "inform" the other in making recommendations. As such, note that the joint model is considerably superior to its $\rho = 0$ analogs in terms of virtually all comparison measures. In all three samples, the homogeneous joint model has lower MAD and RMSE, higher Spearman correlation, and larger mean probability than the analogous (homogeneous) $\rho = 0$ model. Although these increases are modest in some cases (MAD for the joint heterogeneous model is 1.190 compared with the heterogeneous $\rho = 0$ value of 1.262 for the most demanding new-customers/new-movies holdout sample), such improvements to an already sophisticated model are nevertheless nontrivial. This is especially impressive in light of the modest posterior mean of $\tanh^{-1}(\rho)$ of .106 (SE = .042) and relatively large random effect standard deviation, $\sigma_\rho = .235$; we anticipate even greater benefits in applications with more pronounced error correlation.

Joint model compared with alternatives. As we detailed previously, in general, the joint ($\rho \neq 0$) model is superior on all fit measures compared with restricted variants (no covariates, no heterogeneity, no error correlation). The proposed model's improvements over Ansari, Essegaier, and Kohli's (2000) model are in the 10%–20% range (MAD and RMSE) for the new-movies samples, but they are modest for the existing-movies sample. Whereas MAD offers some insight into how much recommended ratings are likely to miss by, the rank-order correlation summarizes a model's capability to prioritize a set of movies correctly in terms of how a particular customer will rate them. That is, the Spearman value accounts for ratings quality in a fundamentally different manner, specifically, whether the model ranks movies similarly to how customers actually do. Note, however, that the Spearman value is driven toward zero by the numerous and unavoidable ties in the discrete rating scale and also by the small set sizes of rated movies; consequently, Spearman values in Table 4 may appear modest in magnitude. Still, the differences between models are large by any standard. Except for the new-customers/existing-movies sample, for which Ansari, Essegaier, and Kohli's model is roughly comparable to the proposed model, differences for the other two samples strongly favor the proposed joint model (which, in one instance, performs best in its homogeneous version). We believe that these differences stem from the additional explanatory power offered by cutoff heterogeneity and, more important, allow for correlation between prediction and selection. Finally, because Ansari, Essegaier, and Kohli's model did not incorporate a selection mechanism, we cannot compare it on this level; however, a relatively simple binary model could readily be incorporated into that model to allow such a comparison.

In short, we find consistent and fairly compelling support for the main modeling constructs introduced at the outset. Taken together, these results demonstrate the importance of accounting for nonignorable missing data in generating

[2]Because MAD treats deviations in any part of the scale identically, it is preferred to squared-error-based measures for ordinal data. We calculated the Spearman correlation in the usual manner, though there are a large number of ties, which tend to lower its value. In addition, when only a single movie was rated (a fairly common event), we assigned a Spearman value of zero, further lowering averaged values for these data. Therefore, the reported values are conservative but equitably so across the models compared. In averaging, Spearman values were weighted by the number of movies in the calculation; this weighting in no way alters the substantive results we report, but it helps correctly reflect customers' influence on the basis of how many movies they rated.

product recommendations. The holdout performance of Ansari, Essegaier, and Kohli's (2000) model is substantially poorer than that for the proposed model. Although raw differences in measures such as MAD and RMSE may appear modest, four points must be mentioned: First, the proportional improvement over Ansari, Essegaier, and Kohli's model is nontrivial; moreover, differences in rank-order correlation are far greater, proportionally speaking. Second, given that the benchmark model is already fairly sophisticated, such an improvement in holdout performance is not a given and is quite satisfactory taken on its own terms. Third, all comparisons to the benchmark model were conditional on ratings being offered, which gives the benchmark the advantage of being compared on its own turf. In actuality, a recommendation consists both of a likelihood of a movie being rated and of the rating itself, and our model is the only one to offer both components. Fourth, recommendation systems are linchpins of online firms proffering billions in recommended merchandise, so that even small performance increases may have major effects on revenue and profitability. Thus, the observed improvements warrant attention from online retailers, which rely daily on product recommendations to drive and expand their business.

DISCUSSION AND CONCLUSION

Our empirical results demonstrate that four modeling constructs—a nonignorable missing-data mechanism, an individual-level account of the ordinal nature of ratings data, a reasonably sophisticated heterogeneity specification, and correlation between the underlying selection and ratings generation processes—can jointly and substantially improve product recommendation accuracy. Although comparisons based purely on model fit statistics do not speak persuasively to practitioners, the degree of improvement in holdout rank correlation does, and it argues in favor of including a comprehensive account of missing data along with a model for product ratings, whatever form it may take. The empirical comparison shows that the full model, including selection and prediction components and heterogeneity, consistently outperforms alternative models in the quality of its recommendations.

The suggested approach provides additional benefits that previous recommendation systems are unable to offer. Chief among these is the partial ability to accommodate customers' selection behavior (i.e., the decision to provide an evaluation for a specific product in the first place). We believe that this feature is of great potential use because company recommendation databases, such as the EachMovie system, include an overwhelming proportion of missing values.

Simply ignoring the missing data and basing recommendations on observed data alone could yield suboptimal or outright inaccurate recommendations. Selection may not be entirely consistent with the customer's rating behavior, as evidenced by our model calibration estimates (Table 2). For example, the customers in our sample are likely to select animated movies and also to give high ratings to them; they exhibit a strikingly different pattern with classic movies, which are selected infrequently but are apparently liked a great deal. Although the proposed models do not pin down the exact causes of these important differences between selection and rating behavior and do not distinguish between possible drivers of selectivity, it may be worthwhile for marketers to investigate the underlying reasons further and act on them.

The knowledge that our model provides can help fashion personalized requests for more ratings input from customers and thus can further improve product recommendations. In other words, our model can not only make more accurate recommendations but also be used to help gather the right customer information to refine the recommendation system in a directed manner. Specifically, the estimated selection model components may serve such a purpose; if a new customer, for whom a recommendation was made based on a high predicted utility for the movie in question, subsequently rates another product, an evaluation may be requested for the recommended (but unrated) product and perhaps other products with a high predicted probability of being selected by that customer. As such, rather than depending on the ratings provided haphazardly by customers or ad hoc rules implemented in the system (as picked up by the selection model), the recommendation system may feed itself with relevant data to improve its own performance. A case in point in our application was that of classic movies, for which the probability of selection was lower than the other genres but the ratings themselves were higher. Ignoring classic movies' low propensity to be selected and focusing merely on them being well liked, given that they are selected, may result in too many positive recommendations for them. Therefore, direct requests for ratings of classic movies may improve the quality of the recommendation system beyond ensuring that a particular sort of movie is suggested only to customers who are likely to appreciate them.

A drawback to building a more complex statistical model for making recommendations—as is the case with our model and those of Ansari, Essegaier, and Kohli (2000) and Chien and George (1999)—is that calibration may be computationally intensive and time consuming. However, calibration of the model needs to be done only at certain time intervals and for restricted samples of customers. Moreover, when the model has been calibrated, the recommendations themselves can be made quickly in real time because the prediction equations are all in closed form. Recommendations for the entire set of samples in our empirical application were generated in several seconds. Such a feature is critical in an online system, in which pages are served up on demand, often many times during each customer visit, as is currently implemented on sites such as Amazon.com. Nonetheless, the model would need to be recalibrated at regular time intervals to reflect evolution in customers' selection and rating of movies. All this is feasible with current technology, so that the modeling framework we developed herein can be readily implemented for existing recommendation systems.

REFERENCES

Allenby, Greg M. and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (1–2), 57–78.

Andrews, Rick L., Asim Ansari, and Imran S. Currim (2002), "Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery," *Journal of Marketing Research*, 39 (February), 87–98.

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), 363–75.

Breese, Jack, David Heckerman, and Carl Kadie (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, WI: Morgan Kaufmann Publisher, 43–52.

Brooks, S.P. and Andrew Gelman (1998), "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7 (4), 434–55.

Canny, John (2002), "Collaborative Filtering with Privacy," paper presented at the IEEE Symposium on Security and Privacy, Oakland, CA (May 2002).

Chien, Yung-Hsin and Edward I. George (1999), "A Bayesian Model for Collaborative Filtering," working paper, Department of MSIS, University of Texas at Austin.

Hofmann, Thomas and Jan Puzicha (1999), "Latent Class Models for Collaborative Filtering," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Madison, WI: Morgan Kaufmann Publisher, 688–93.

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

Rost, Jurgen (1985), "A Latent Class Model for Rating Data," *Psychometrika*, 50 (1), 37–49.

Rubin, Donald B. (1976), "Inference and Missing Data (with Discussion)," *Biometrika*, 63 (3), 581–92.

Sarwar, Badrul M., George Karypis, Joseph A. Konstan, and John T. Riedl (2000), "Analysis of Recommendation Algorithms for E-Commerce," working paper, GroupLens Research Group/Army HPC Research Center, University of Minnesota.

Wedel, Michel, Wagner A. Kamakura, Albert C. Bemmaor, J. Chiang, Terry Elrod, R. Johnson, Peter J. Lenk, Scott A. Neslin, and C.S. Poulsen (1999), "Discrete and Continuous Representation of Heterogeneity," *Marketing Letters*, 10 (3), 217–30.