

Offering Online Recommendations with Minimum Customer Input Through Conjoint-Based Decision Aids

Arnaud De Bruyn

Department of Marketing, ESSEC Business School, 95000 Cergy, France, debruyn@essec.fr

John C. Liechty

Smeal College of Business, The Pennsylvania State University, University Park, Pennsylvania 16802, jcl12@psu.edu

Eelko K. R. E. Huizingh

Department of Business Development, University of Groningen, 9700 AV Groningen, The Netherlands, k.r.e.huizingh@rug.nl

Gary L. Lilien

Smeal College of Business, The Pennsylvania State University, University Park, Pennsylvania 16802, glilien@psu.edu

In their purchase decisions, online customers seek to improve decision quality while limiting search efforts. In practice, many merchants have understood the importance of helping customers in the decision-making process and provide online decision aids to their visitors. In this paper, we show how preference models which are common in conjoint analysis can be leveraged to design a questionnaire-based decision aid that elicits customers' preferences based on simple demographics, product usage, and self-reported preference questions. Such a system can offer relevant recommendations quickly and with minimal customer input. We compare three algorithms—cluster classification, Bayesian treed regression, and stepwise componential regression—to develop an optimal sequence of questions and predict online visitors' preferences. In an empirical study, stepwise componential regression, relying on many fewer and easier-to-answer questions, achieved predictive accuracy equivalent to a traditional conjoint approach.

Key words: conjoint analysis; recommender system; online decision aid; efficiency

History: This paper was received December 14, 2006, and was with the authors 27 days for 1 revision.

Published online in *Articles in Advance* March 31, 2008.

1. Introduction

Customers search and process information in order to choose which product or service to buy from the many available options. Researchers have long acknowledged that when searching, customers do not necessarily attempt to find the optimal solution (Wright 1975). Due to the number of alternatives, the complexity of comparing alternatives on multiple dimensions, time pressure, customers' limited information processing capabilities, and cognitive efforts associated with decision making, customers rarely process all available information when choosing an alternative: The incremental utility of a better-than-good alternative might not justify the additional time and "thinking costs" necessitated by the task (Shugan 1980). In other words, "consumers may often have to compromise between optimizing eventual consumption benefits and reducing the strains of decision making" (Wright 1975, p. 62). Hence, identifying satisfactory alternatives may suffice—a process referred to as "satisficing" (Simon 1957).

On the Internet, where customers are known to be impatient (Banister 2003) and where available information is often overabundant, the trade-off between search costs and decision quality might very well be exacerbated. It is therefore crucial for e-commerce Web sites to recognize this problem and to offer recommender systems to help their visitors search their catalogs of offers more efficiently.

In their attempt to help customers make their decision process as efficient as possible, researchers and online merchants have developed a variety of strategies (such as featuring stores in e-marketplaces; see He and Chen 2006) and online decision aids (such as collaborative filtering, information filtering, decision-support systems, etc.; see next section). Consider these two recent and well-publicized examples: (1) Netflix (2006) announced a \$1,000,000 prize to researchers who could substantially improve the accuracy of their movie recommendation system; (2) Wal-Mart had to publicly apologize after its retail Web site made offending movie suggestions such as recommending

DVDs with an African American theme to consumers browsing “Planet of the Apes” DVDs (*The Washington Post* 2006).

This paper investigates a particular type of system: *questionnaire-based decision aids*, whose purpose is to elicit customers’ preferences through a sequence of easy-to-answer questions while requiring little prior information. Currently this approach is used mainly to screen out alternatives rather than to fully assess customers’ preferences. Our goal here is to compare alternative ways of embedding a preference model, such as those used in conjoint analysis, in questionnaire-based decision aids in order to *weight* and *rank* rather than *screen* alternatives. Our work is in line with that in the computer science literature to incorporate preference models in multiattribute decision-making recommender systems (Choi et al. 2006, Weng and Liu 2004). The computer science work has not incorporated the insights from the literature on conjoint analysis, preference models, and preference elicitation procedures, and falls short of addressing how consumers actually make purchasing decisions. In this paper, we leverage the ability of conjoint analysis to represent customers’ preferences in building an efficient questionnaire-based recommender system, while avoiding both the complexity and the traditional customer burden associated with lengthy conjoint analysis data collection. To be useful, we seek a system that satisfies the following requirements:

- Customer input should be minimal.
- The method should not require prior knowledge about the customer.
- Product-category expertise should not be needed to use the recommender system.

In the next section, we review the available types of recommender systems and discuss how questionnaire-based methods complement existing ones. Then, we discuss how conjoint analysis can be used to build more efficient questionnaires and present three competing methods that use the results of a conjoint study conducted *ex ante* to develop an optimal sequence of questions to elicit customers’ preferences. We conduct an empirical test of the methods and find that *stepwise componential segmentation* elicits customers’ preferences and makes quality recommendations that compare favorably with those from a full profile conjoint study after respondents answered only two simple questions. We conclude by discussing the results and their theoretical and managerial implications.

2. Recommender Systems

Various types of online recommender systems have been developed, and that variety reflects the heterogeneity of online visitors’ needs, search preferences,

and capabilities as well as the information available to the Web site about its visitors that can be used to make recommendations.

Collaborative filtering recommender systems (Resnick and Varian 1997, Schafer et al. 2001) are “agents that use behavioral or preference information to filter alternatives and make suggestions to a user” (Ansari et al. 2000). Collaborative filtering draws on a database of customers’ ratings, preferences, past purchases, or browsing behavior to predict a visitor’s affinity for items based on comparisons to other customers with similar tastes (Konstan et al. 1997, Resnick and Varian 1997, Schafer et al. 2001, Shardanand and Maes 1995). Typically, when a customer browses a book at Amazon.com, a collaborative filtering system parses the company’s database to predict what other books that customer might be interested in (i.e., “Customers who bought this book also bought...”). This approach dynamically adapts the recommendations it makes to the preferences of existing customers as revealed through their actual purchases. However, to be efficient, such systems require some information about visitors’ preferences and are therefore not well-suited for first-time visitors. In addition, collaborative filtering often “provides few, if any, reasons for a recommendation” (Ansari et al. 2000, p. 363), and many collaborative filtering algorithms act mainly as a “black box” (Herlocker et al. 2000). They are also notoriously sensitive to information scarcity (Popescul et al. 2001) and can lead to unwise recommendations in the absence of a sufficiently large comparison base.

Customer decision support systems (CDSS) are systems that “connect a company to its existing or potential customers, providing support for some part of the customer decision-making process” (O’Keefe and McEachern 1998). Various screening tools based on self-reported preferences and comparison matrices fall into that category (Haubl and Trifts 2000). Early developers assumed that by facilitating the manipulation of information and expanding their information processing capabilities, users of CDSS were likely to compare more alternatives, evaluate them more completely, and thus make better decisions (Hoch and Schkade 1996). Although such predictions have not always been confirmed, the use of CDSS in online environments increases search efficiency and choice quality (Haubl and Trifts 2000). However, most CDSS assume that customers are willing and capable of comparing alternatives on those performance dimensions that are relevant and important to them. This assumption is questionable with complex, intangible, or highly customizable products and where customers’ expertise is low (Grenci and Todd 2002, Huffman and Kahn 1998) or when potential customers are impatient and unwilling to go through

a time-consuming evaluation procedure. Under such circumstances, it can be risky for a commercial Web site to put such a burden on its visitors.

Questionnaire-based decision aids seek to mimic the interactions between customers and sales representatives encountered in traditional brick-and-mortar sales environments and are especially well-suited for multiattribute decision-making products. Online visitors' preferences are elicited through a series of questions and the decision aid offers product or service recommendations accordingly. For instance, "a carshopper may need to provide answers to such questions as what type, size, and features she prefers, what price she can afford, whether luxury or economy (lower cost, better fuel, etc.) is more important to her. The system then searches its knowledge base for cars that best satisfy these requirements" (Tran 2006, p. 2). Such a system can recommend brand new or rarely purchased products as long as they fit customer interests (Weng and Liu 2004), while collaborative filtering requires a comparison base of past purchases to recommend items with confidence. DeLong et al. (2005) emphasize the need to design efficient, dynamic questionnaires to identify the next most informative question to ask, and acknowledge that nonexpert users and new visitors are especially challenging to handle due to the large knowledge gap that separates customers' needs and firms' offers.

Bergmann et al. (2002) observe that most implementations are in the form of static and rather technical questionnaires, and draw on direct preference elicitation procedures (e.g., product specifications) rather than indirect questions (e.g., customer's profile, intended product usage). Such systems are also referred to as *rule-based information filtering systems* (Kuflik et al. 2003), *parametric search engines* (Kamis and Stohr 2006) or *parameter-based interfaces* as opposed to *needs-based interfaces* (Randall et al. 2005). "[Such questionnaires] are a convenient way for experts to express their product wish in every detail but can be very hard to handle for inexperienced customers" (Bergmann et al. 2002, p. 8). One exception is MyProductAdvisor (<http://www.myproductadvisor.com>) which, in addition to technical requirement questions, asks a few intended product usage questions (e.g., "I plan to use my new computer often when I am traveling"). Recommendations based on these intended product usage questions can be unsatisfactory.¹

¹ E.g., for laptop computers, the system asks six product intended questions. We portrayed two radically different users; a mostly sedentary, light user ("Applications for home/business and Web browsing"), and a heavy user of media technologies ("Playing digital media," "Editing digital media," "Playing the latest games"). Based solely on this information, MyProductAdvisor recommended

Questionnaire-based decision aids usually involve option-screening rather than option-weighting mechanisms. With an *elimination-by-aspects* approach, the user "narrows the set of alternatives, one attribute at a time" (Kamis and Stohr 2006). With *parametric search*, the user imposes upper and lower bounds for one or more attributes while ignoring others (Hagen et al. 1999). In any case, alternatives are sequentially eliminated based on customers' answers (stated preferences or constraints) until only a few options remain. In the process, customers' stated constraints might become overly restrictive and the decision aid might not be able to find products that satisfy them all, resulting in an empty recommendation set. In the absence of a customer's preference model, the decision aid can not determine which constraints could be relaxed with minimal impact on customer satisfaction. The most common solution is then to ask the customer to relax constraints until a nonempty recommendation set is generated. For instance, a restaurant recommendation system might not be able to satisfy a customer's request for a "fancy Italian restaurant within a 10 mile radius" (Johnston et al. 2001). If the closest available alternatives were a casual Italian restaurant nearby, a fancy French restaurant at 8 miles, and a fancy Italian restaurant at 11 miles, the decision aid has no way of knowing which to suggest. One approach is to relax constraints on neighborhood specifications first (e.g., expand geographic limit from 10 to 15 miles), then relax whatever constraint leads to the smallest set of alternatives (Chung 2004). In the absence of a customer preference model, it is not clear that the constraint chosen to be relaxed will lead to the best recommendation. Given the difficulties such decision aids pose, merchant Web sites such as Lycos Shopping, MSN Shopping, Dealtime/Shopping.com, and MySimon have abandoned this type of approach (Kamis and Stohr 2006).

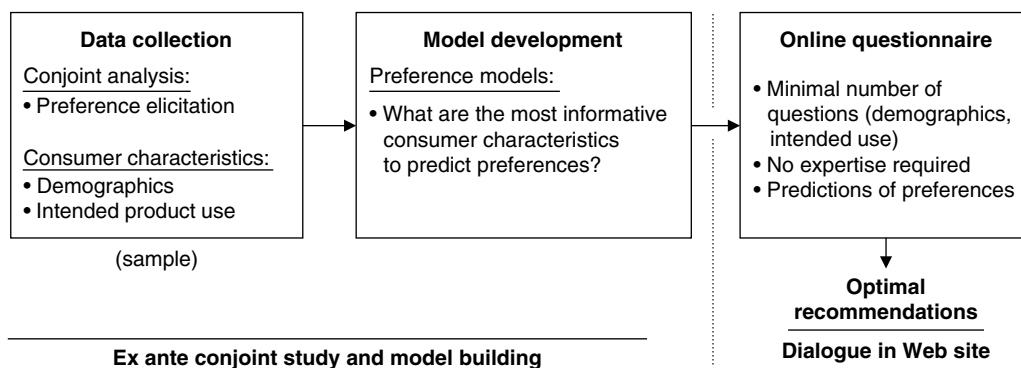
In the next section, we discuss how to build on conjoint analysis preference models to design more efficient and informative questionnaire-based decision aids. This approach follows Ansari et al.'s (2000) observation that preference models used in marketing offer good alternatives to collaborative and information filtering recommender systems when prior behavioral data about an individual is sparse.

3. Methodology

Despite efforts to make preference elicitation procedures quicker and more efficient (Sawtooth Software

five computers to each user; four of which were recommended in both cases (80% of recommendation overlap, despite radically different user profiles). These recommendations produced a mismatch between expressed needs and computers' performance. This lack of personalization highlights the potential for improvement in such systems.

Figure 1 Conjoint-Based Recommender System. We Propose a Three-Step Approach to Develop an Optimal Sequence of Questions to Elicit Online Visitors' Preferences and Make Optimal Recommendations Based on a Conjoint Study Conducted Beforehand



2002, Toubia et al. 2003), conjoint analysis still requires considerable customer input and is usually impractical as the core of a recommender system. Nevertheless, conjoint analysis offers a useful way to model customers' preferences. Hence, we explore the possibility that a conjoint study conducted beforehand on a sample of customers can be leveraged to design a recommender system capable of offering personal recommendations to online visitors with minimal inputs. We employ the following approach (see Figure 1):

- We perform a conjoint analysis on a representative sample of individuals who, in addition to selecting, ranking, or rating a set of products, are also asked to answer demographic, product usage, and self-reported preference questions.
- We link respondents' characteristics to their preferences and we identify the most informative demographic and product usage questions.
- We use the results of this analysis to develop an optimal sequence of questions to elicit customers' preferences and make recommendations.

We now focus on the second stage of this approach—identifying the most efficient questions to ask in order to elicit a customer's preferences—and discuss three competing methods to link individuals' responses to their preferences in a way that can be operationalized in an online questionnaire.

3.1. Cluster Classification

A natural approach is to follow a three-step, segmentation-targeting-like strategy similar to those commonly implemented in direct marketing. After conducting the conjoint study and collecting individual-level responses (e.g., ratings, preferred choices, or pairwise comparisons of conjoint profiles), (i) individuals' preference partworths are estimated with standard estimation procedures; (ii) respondents are then clustered into segments of similar needs and preferences, using either hierarchical or nonhierarchical methods (Green and Krieger 1991); and (iii) descriptor variables or segmentation bases (e.g., demographics

characteristics, intended product usage, self-reported preferences) are used to predict segment membership of each individual. We refer to these three steps as the *estimation*, *clustering*, and *classification* stages, respectively.

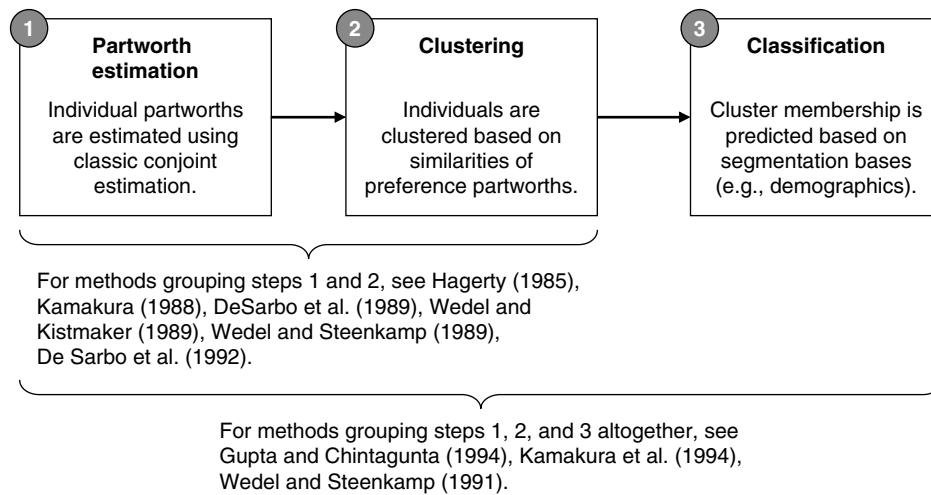
Within the context of a recommender system, the classification procedure identifies the relevant questions to ask in order to determine to which segment an online visitor is most likely to belong. That predicted segment membership is exploited to make those recommendations best-suited for a typical member of the identified segment.²

Although the *estimation*, *clustering*, and *classification* procedures are usually envisaged separately, both in academic (Green and Krieger 1991) and commercial applications (Wittink et al. 1994), the combined approach has the shortcoming of grouping respondents based on possibly unreliable individual-level estimates (the degrees of freedom at the individual level being usually rather small). In addition, conjoint models that are overparameterized at the individual level can not be accommodated; respondents must express their preferences (e.g., rating of choice task) for at least as many profiles as there are attribute levels to be estimated. Finally, segments are formed regardless of their actual targetability, and the *clustering* and *classification* stages might achieve suboptimal solutions because they try to maximize two separate objective functions independently.

It has been suggested that traditional conjoint-based segmentation could be improved by grouping

²This approach shares many similarities with the work of Kamakura and Wedel (1995; see also Balasubramanian and Kamakura 1989, Singh et al. 1990), where (a) they administered a lengthy lifestyle questionnaire, (b) found latent clusters in the population of respondents, and (c) identified in the initial questionnaire the most informative questions to predict cluster membership, hence reducing the data collection by 78% while recovering 73% of the information. For our type of application, however, two different data sets are involved: one to define and identify the clusters (preference partworths from conjoint data) and another to assign segment membership (questionnaire).

Figure 2 Traditional Conjoint-Based Segmentation Usually Follows a Three-Stage Procedure



Notes. The three-stage procedure is as follows: (i) Individual preference partworths are estimated for each respondent. (ii) Individuals are grouped into homogeneous segments based on preference partworth similarities (e.g., K means). (iii) Segment membership is predicted based on available descriptors (e.g., discriminant analysis). Researchers have argued that grouping two or more steps together could increase overall performance.

two or more stages together. Figure 2 summarizes these developments.

Researchers have developed various methods to group the *estimation* and *clustering* stages into a one-step procedure that optimizes a single objective function. Such methods include Q-factor analysis (Hagerty 1985), hierarchical clustering (Kamakura 1988), clusterwise regression (DeSarbo et al. 1989, Wedel and Kistmaker 1989, Wedel and Steenkamp 1989) and mixture regression methods (DeSarbo et al. 1992). A Monté Carlo study (Vriens et al. 1996) showed that, in out-of-sample predictive accuracy, none of these methods outperformed the traditional, two-step approach, which consists of separately estimating individuals' partworths and segmenting the population. This finding can be explained by the within-segment heterogeneity that affects all methods and their resulting performance similarly (Wedel and Kamakura 2000). Therefore, although the above methods differ greatly on other performance criteria (Vriens et al. 1996), the actual method employed does not significantly affect the out-of-sample predictive accuracy of preference partworths based on segment membership.

Other researchers have proposed methodologies to group all three stages, *estimation*, *clustering*, and *classification*, into an integrated framework (Gupta and Chintagunta 1994, Kamakura et al. 1994, Wedel and Steenkamp 1991). While these algorithms have merits, they are not suited for the application we address. Because they draw *simultaneously* (as opposed to *sequentially*) on all available descriptors to assign segment membership, these methods can not identify the most informative questions to ask nor indicate the

order in which questions should be asked to best predict segment membership.

The same problem affects traditional classification methods such as discriminant analysis or artificial neural networks. These methods use all information available simultaneously and therefore can not identify a sequence of questions to predict the most likely cluster of preferences for an individual. A variant of discriminant analysis, stepwise discriminant analysis, would seem to offer a potential solution but the method has been criticized in the literature for poor out-of-sample prediction (Huberty 1994, 1989).

The CART (classification and regression trees) algorithm offers an interesting alternative (Breiman et al. 1984). CART sequentially splits a population (the parent node) into child nodes, such that each child node is populated with individuals as pure as possible in terms of class membership. Then, each child node becomes a parent node itself and is subsequently split and so on, until a stopping rule is reached. The tree is eventually pruned back based on a cost-complexity criterion to reduce overfitting and enhance out-of-sample predictive accuracy. Ideally, each end node of the tree becomes perfectly pure; it contains individuals from only one segment, achieving a perfect classification.

CART has several useful properties for the focal application. First, the solution tree structure can readily be translated into an optimal sequence of questions, each split pointing out the next best additional bit of information. Second, two child nodes could be expanded using two different splitting rules (i.e., subsequent best splits might differ in the left and right child nodes); hence, splits are locally optimal. Finally, it is straightforward to make recommendations based

on the tree's structure: preferences of the segments populating each node can be used directly for an out-of-sample population.

For this research, we integrate the above methods into the traditional, three-step approach as follows:

Estimation. Individual-level preference partworths are first estimated using classical conjoint equations.

Clustering. Individual partworth estimates are then clustered into preference segments. We favor a non-hierarchical clustering methodology (i.e., K means) because centroids have an immediate interpretation representing the average preference partworths of the segment's population and can be readily translated into optimal recommendations.

Classification. Cluster membership is predicted based on descriptor variables using CART. The tree structure conveys both the sequence of questions to ask (i.e., the sequence of descriptor variables employed by CART to split the population) and the optimal recommendations to make (i.e., average preference partworths of each node's population).

This integrated approach also has the advantage of being easily replicable with standard statistical packages, hence being more readily implementable.

Several researchers have claimed that segmenting and pooling similar individuals could improve predictions for each individual in conjoint analysis (DeSarbo et al. 2002, 1989, 1992; Green et al. 1993; Kamakura 1988; Ogawa 1987). Pooling individuals increases the degrees of freedom of the model, leads to more stable and accurate partworth estimates, and prevents the model from overfitting individual-level data. Bayesian treed regression and stepwise componential segmentation follow this line of thinking: They both integrate *estimation*, *clustering*, and *classification*, and pool data obtained from similar individuals. However, in contrast to the single stage methods cited earlier, classification is achieved separately and hence can be used to identify an optimal sequence of questions.

3.2. Bayesian Treed Regression

The idea behind *Bayesian treed regression* is to partition a data set using a tree structure but instead of computing a simple mean or proportion, fitting a different regression model at each end node (Chipman et al. 2002). Essentially, Bayesian treed regression combines CART and clusterwise regression within a hierarchical Bayes regression framework.

Although Chipman et al. (2002) did not develop the algorithm with conjoint analysis in mind, its application to this domain is straightforward. The treed regression algorithm simultaneously (i) clusters individuals into nonoverlapping segments (i.e., nodes) through a tree structure, (ii) pools profile ratings made by individuals in the same node into a unique regression model, and (iii) estimates preference partworths

using hierarchical Bayes regression. Because parameters of the conjoint model are computed at the node level, treed regression avoids the overfitting issue of individual-level models (unless a single individual populates a node). Furthermore, segments are formed on the basis of a sequence of binary splits performed on descriptor variables. The segments are therefore perfectly identifiable and, as in CART, the optimal sequence of questions is naturally embedded in the solution.

In contrast to traditional tree methods that apply locally optimal, greedy splitting rules, Bayesian treed regression tries to achieve global optimality by searching the space of possible trees using Markov chain Monte Carlo (MCMC) exploration (Chipman et al. 2002). The MCMC algorithm samples from the posterior density of possible trees and results in a stochastic search of the solution space, where most of the search is concentrated on parts of the space that have a high probability conditional on the data and prior specification.

3.3. Stepwise Componential Segmentation

While both CART and Bayesian treed regression are natural candidates for the design of an optimal sequence of questions, they may be subject to overfitting because each split of the tree retains a smaller portion of the data set. In addition, data requirements grow exponentially with tree size.

Several hybrid methods have been proposed to overcome the heavy data requirements of classic conjoint analysis and to reduce data collection effort and time (Green 1984). One of these approaches, *componential segmentation* (Green and DeSarbo 1979), explicitly incorporates respondent descriptor variables in the utility function by reexpressing individuals' preference partworths as linear combinations of descriptor variables.

We use the following notation:

- 1.. i .. I refers to respondents.
- 1.. j .. J refers to profiles rated by each respondent.
- 1.. k .. K refers to preference partworths to be estimated by the model, i.e., one per attribute level (excluding all dummy levels set to zero for identification purpose), including an intercept.
- 1.. q .. Q refers to respondents' descriptor variables such as demographics, intended product usage, etc.
- y_{ij} is the preference score given by the i th individual to the j th profile.
- β_i is the vector of preference partworth of the i th individual (K elements).
- X_{ij} is the vector of attribute levels of the j th profile rated by the i th individual (K elements).

- D_i is the vector of descriptor variables pertaining to the i th individual (Q elements).
- Ψ is a matrix of parameters to be estimated (K rows and Q columns). This matrix is not specific to a particular individual.

We build a traditional conjoint model where predicted preference scores \tilde{y} are linear combinations of preference partworths and attribute levels, $\tilde{y}_{ij} = (\beta_i \cdot X_{ij})$, such that they minimize:

$$SSE = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \tilde{y}_{ij})^2. \quad (1)$$

Whereas classic conjoint model computes all vectors β_i individually, we reexpress:

$$\beta_i = (\psi \cdot D_i): \forall i \quad (2)$$

and optimize Ψ at the population level. In other words, we approximate individuals' preference partworths as a linear function of individuals' descriptors (D_i , a vector of independent variables) and a matrix of parameters (Ψ) to be estimated. Ψ elicits and quantifies the statistical relationships between individuals' descriptors and preferences, and is estimated at the population level. While in traditional conjoint estimation there are IK parameters to be estimated, there are QK parameters in componential segmentation, with Q commonly much smaller than I , hence increasing the degrees of freedom for estimation.

Although this approach seems natural (one should expect price sensitivity to be a function of income or preferences for specific benefits to be linked to demographics, lifestyle characteristics, or intended product usage), componential segmentation has had a limited impact in practice for at least two reasons. First, componential segmentation "leads only to subgroup utility functions because all respondents with a similar background profile are assumed to have the same utility function" (Green 1984, p. 156). In other words, if $D_i = D_j$, then $\beta_i = \beta_j$: respondents with identical descriptors are assumed to have identical preferences.

Second, the success of this method depends on existing correlations between consumers' characteristics and individual preferences (Wedel and Kamakura 2000). If the latter are only loosely related to observable respondents' characteristics, results will be weak. Given the nature and limited amount of information usually available in segmentation applications, this requirement can be a serious impediment.

This limitation, however, is much less critical in a recommender system context because very specific questions can be asked to online visitors, including those pertaining to normally unobservable consumers' characteristics (i.e., needs, likes and dislikes,

experience, intended product usage), broadening the range of possible and relevant questions.

In its original format, all information is included in the componential segmentation model; preference partworths' estimation draws on all Q individuals' descriptors. In order to determine an optimal sequence of questions, we need to adapt the original algorithm to make it a stepwise procedure as follows.

We assume that $Q = 1$ (vector of descriptors is of size 1) and set $D_i = \{1\}: \forall i$ (equivalent to an intercept in linear regression). Hence, Ψ is a vector of size K and $\beta_i = \Psi: \forall i$. Ψ is equivalent to the average preference partworths of the population as a whole, which minimizes SSE .

The vectors of descriptor variables are then augmented by one element ($Q' \leftarrow Q + 1$) at each step of the stepwise procedure. The next descriptor included in the model is the one that minimizes SSE , conditional on the optimization of the new matrix Ψ . As with the CART algorithm, the selection of the next most informative descriptor is achieved by testing all possible descriptors one by one³ as potential candidates to fill the last element of D , and by eventually adding the one to the descriptor matrix that leads to the highest incremental improvement. At each step, the matrix Ψ is augmented by one column, as is the number of parameters to be estimated by K elements.

A "statistically optimal" stopping rule would be to test the hypothesis that the last K parameters added to the model are not different from zero. If this is true, the new model is no better than the previous one with K fewer parameters; the last descriptor is removed and the stepwise procedure stops. This hypothesis can be tested with an F test, distributed as $F_{K, I, J-K(Q-1)}$ (Rencher 1995, p. 359).

In practice, however, the large number of degrees of freedom will make the rejection of the null hypothesis very unlikely (IJ represents the number of profiles rated by the total number of respondents, several thousands in most conjoint studies). Translated into the context of recommender systems, if this stopping rule were to be applied, the questionnaire suggested by this method would be too lengthy. The recommender system would keep asking for additional information as long as it would make statistical sense, even if improvements in actual recommendations were marginal.

³Notice that a Likert question, for instance, can be operationalized in different ways. The raw scores can be inserted directly in the descriptor matrix, thus resulting in a linear structure. Alternatively, the population can be described in terms of binary splits (e.g., for a Likert 7 question, those who answered 2 or less and those who answered 3 or more), allowing both linear and nonlinear relationships between consumers' characteristics and preference partworths to be used as predictors.

Therefore, we chose a more practical stopping rule, stopping questionnaire development when the inclusion of an additional descriptor does not improve the adjusted R^2 (between y and \hat{y}) by at least 0.005. We favor adjusted R^2 as a performance metric to take into account the increasing number of parameters at each step.

In contrast to the optimal sequence of questions suggested by the two previous tree-based methods, stepwise componential regression generates static questionnaires: Online visitors get the same questions in the same order, independently of their answers to previous questions. However, each question selection is optimized on the entire data set, which is likely to enhance the robustness of the method and reduce the risk of overfitting.

In the next section, we report the results of an empirical study in which we tested the three algorithms described here, *cluster classification*, *Bayesian treed regression*, and *stepwise componential segmentation*, as means to build an effective recommender system. Because this research addresses recommendation quality and efficiency, we use out-of-sample predictive ability compared to the length of the questionnaire suggested by each method as our critical measure of performance effectiveness.

4. Empirical Assessment

4.1. Research Design

We recruited 616 graduate and undergraduate students at a large northeastern U.S. university to rate customized Web pages from a hypothetical university news portal. We conducted the study electronically in a controlled lab setting, and the pages to be rated were displayed on the participants' computer screen. The pages were described by five attributes: weather report, university-related news, general news, business news, and value of an online coupon. Attributes had either two or three levels. Table 1 reports the five attributes and their levels, as well as the estimated average preferences in terms of preferred levels⁴ and total variance explained, estimated using individual-level conjoint models.

The study comprised four parts. In the first part, participants were familiarized with the attributes and the software system used to collect the data.

⁴A level is deemed to be "preferred" as soon as its preference score is higher than the preference scores of the other levels, even if this difference is not statistically significant. In the case of the online coupon, about half of the respondents were mostly indifferent (online coupon captures only 12% of the variance), some with very low positive values, others with negative values, which explains why 26% of the respondents are reported to "prefer" the online coupon with the lowest dollar value.

Table 1 Attributes and Attribute Levels of the Research Design

Attribute	Levels	Preferred level (%)	Variance explained (%)
Weather report	5-day forecast	60	15
	1-day extended report	40	
University news	General news	51	17
	Sports news	49	
Online coupon	\$2.00	26	12
	\$4.00	74	
General news	U.S. news only	11	29
	Mix U.S./world	76	
	World news only	13	
Business news	Stocks news only	8	27
	Mix stocks/general	58	
	General news only	34	

Notes. Most respondents prefer a 5-day weather report, general university news, a \$4.00 online coupon, and a mix of both general and business news. The most important attribute in explaining preferences for particular pages is the general news attribute.

The second part of the study was a *conjoint task*, where respondents rated 21 Web pages, displayed one at a time on the screen, on a 100-point preference scale. To select the 21 profiles shown to participants during the main conjoint task, we built 4 partially balanced blocks using an orthogonal fractional factorial design, so that news pages evenly spanned the space of possible combinations and attribute levels were equally represented. Each participant was randomly assigned to one of these blocks. The order of the profiles to be rated was randomly rotated within each block and across respondents.

The third part of the study was a self-administered *questionnaire*, with 99 questions on respondents' socio-demographics, consumption habits, likes, and dislikes. The questions were selected following a pilot study conducted with 43 students with similar backgrounds to the study sample, who were asked to explain the rationale behind their preferences for certain attributes. Additional questions were also suggested by experts. For instance, one of the questions retained was whether or not participants owned common stock, a likely influence on their preferences for business news. Of the 99 questions, 44 were dichotomous, 7 were multiple-choice questions, and 48 were Likert-scaled items.

The final part of the study was a *holdout exercise*, where respondents distributed 100 points amongst 4 different news pages. We used this part of the study to assess the out-of-sample accuracy of the recommender systems. We built 20 different news pages using a fractional factorial design and grouped them into 5 balanced sets of 4 pages each. We showed these 5 sets to participants, one at a time (i.e., on five different screens), and asked them to distribute 100 points amongst the 4 different news pages

displayed on the screen, with more points indicating stronger preferences. To increase comparability, we showed all participants the same five sets of pages but randomly rotated the order to avoid presentation effects.

Note that because the chosen product category (news pages) does not lead to natural cross-attribute trade-offs, a compensatory preference model such as conjoint analysis would have little added value in a real-life setting: An online merchant could simply ask a few direct questions designed to help them create their ideal personal news page. Our holdout exercise, however, presents a few limited options requiring respondents to exert trade-offs among attributes; the ability to measure and predict respondents' trade-offs is the essential ingredient in predicting choices, our key research goal and one thus satisfied by our setting.

4.2. Analysis

We performed the analysis using tenfold cross-validation. We randomly split the data into 10 training sets ($N = 554$, 90% of the sample) and 10 corresponding testing sets ($N = 62$, 10%), and ensured that all individuals would appear in one and only one testing set. We performed data analyses on the training sets only, that is, we based the sequences of questions suggested by all three algorithms to make recommendations exclusively on information available for the 10 training sets of 554 individuals each. The respondents in the testing sets were those to whom recommendations were made. We used the results of the holdout task of these individuals to assess the quality of the simulated recommendations. We report parameter estimates and other results in this section from the first cross-validation sample for clarity and consistency; the performance assessment results we report in §4.4 are averaged over all ten cross-validation samples.

For the cluster-classification algorithm, after estimating and scaling the individuals' partworths, we grouped respondents into clusters of preferences using the K means algorithm. We repeated each K means analysis five times with different starting values to reduce the risk of local minima, and chose the optimal number of clusters using silhouette analysis (Kaufman and Rousseeuw 1990). We found that 6 groups worked best for 8 of the 10 cross-validation samples, while a 4 and a 5 segment solution worked best for the other two, respectively. Sensitivity analysis showed that increasing the number of clusters did not significantly improve the out-of-sample fit for this method. Table 2 reports the size and average preference partworths of the six identified clusters (numbered C1 to C6) for the first cross-validation sample.

We built the classification tree using the CART algorithm and stopped the splitting process when

Table 2 Average Preference Partworths of the Six Identified the Clusters of Respondents, Labeled C1 Through C6 (First Cross-Validation Sample)

Attribute	Levels	C1	C2	C3	C4	C5	C6
		$N = 114$	$N = 99$	$N = 79$	$N = 63$	$N = 79$	$N = 121$
Intercept		46.1	15.5	49.9	21.3	75.1	51.0
Weather report	5-day forecast (*)						
	1-day extended report	-27.0	-1.6	-3.8	20.2	-16.0	-2.7
University news	General news (*)						
	Sports news	10.8	-0.9	-26.0	16.9	-1.2	-13.8
Online coupon	\$2.00 (*)						
	\$4.00	3.0	9.8	6.4	7.7	-2.6	-0.3
General news	U.S. news only (*)						
	Mix U.S./world	10.7	29.6	22.0	11.3	-9.2	4.0
	World news only	-5.8	10.3	9.6	-8.9	-26.2	-22.3
Business news	Stocks news only (*)						
	Mix stocks/general	17.5	29.0	-2.5	7.9	-2.6	27.9
	General news only	4.0	17.6	3.8	6.2	-2.6	19.1

Notes. A typical member of the fourth cluster (C4) highly values news and extended weather forecast report but does not care much about the type of business news displayed or the face value of the online coupon. All types of general news are okay as long as they contain U.S. news, although a mix is preferred. Note that a mix of U.S. and world general news dominates the other general news options for five of these six clusters.

(*) Dummy levels set to zero.

we reached a minimum node size, and then pruned it back using the cost-complexity criterion (Breiman et al. 1984). The final tree had an average depth of 4.0 splits and a maximum depth of 11 splits.

For the Bayesian treed algorithm, given our objective to require minimal customer input, we set the four parameters that govern the splitting decisions to values that favor small trees ($\alpha = 0.5$, $\beta = 2$, $c = 1$, and $\lambda = 0.404$; see Chipman et al. 2002 for discussion). We ran a sufficiently large number of iterations to assure the stability of the solution and dedicated one-third of the training set to develop an internal overfitting diagnostic. The final tree had an average depth of 4.2 and a maximum depth of 6 splits.

The stepwise componential regression did not require any parameterization, and tests of the modifications in the adjusted R^2 led us to stop the development of the model after the second question for all tenfold cross-validations.

4.3. Out-of-Sample Test Design

We calculated reference partworths for the individuals retained for out-of-sample testing (10% of the population for each cross-validation) based on the analysis of the training sample.

For the cluster-classification method, we initially set probabilities of belonging to each of the clusters equal to the proportion of that cluster in the in-sample population. Then, individuals in the testing set navigated the estimated tree based on their answers to the demographic and product usage questions, and probabilities of their cluster membership were updated

Table 3 Example of Questionnaire Suggested by the Cluster-Classification Method and Associated Responses from the 53rd Respondent

Question	Answer	Probability to belong to cluster 4 (%)
Initial proportion (size of the cluster in parent node)		11.4
Q1. General [university] news is typically more important to me than [university] sports news.	“No”	21.7
Q2. With regards to local weather reports, detailed summaries of today’s weather is typically more important to me than a less detailed five-day forecast.	“Agree”	39.3
Q3. In the last month, how many times have you eaten at a restaurant (do not include on campus restaurants)?	“3 to 6 times”	50.0
Q4. Have you ever taken a seminar or class about the Web or Internet?	“No”	63.3
Q5. How many men’s home basketball games have you attended so far this season?	“3”	85.7

Notes. Prior to the first question, the probability of an individual to belong to the 4th cluster is equal to the proportion of this cluster in the sample, i.e., 11.4% (63 respondents Out of 555). After 5 questions, the system estimates that this individual has an 85.7% chance to belong to the 4th cluster and stops asking additional questions. Recommendations are then optimized based on the prediction of cluster membership. Each question corresponds to a node in the tree developed by the CART algorithm.

after they answered each question. At each step, their estimated preference partworths were an appropriately weighted average of the preference partworths in their assigned cluster. Table 3 reports an example of how one individual answered the questionnaire and how his probability of belonging to the fourth cluster was updated based on his answers.

We applied the same procedure for the Bayesian treed regression method: Individuals navigated the estimated tree and their preference partworths were updated based on their answers. The Bayesian treed regression provides a different set of estimates for each end node, making it simple to assign preference partworths to individuals when they reach a final node. For nonfinal nodes, because Bayesian treed regression optimizes the tree globally and does not provide intermediate results, we computed intermediary partworths as a weighted average of the partworths in the remaining, downward nodes of the tree.

For the stepwise componential segmentation approach, partworths are a direct function of the questions answered. Table 4 indicates how preference estimates are updated after the first three questions, i.e., the table reports Ψ , the parameter estimates, at each step of the development of the questionnaire.

Before respondents are asked any questions, Ψ is a vector of eight elements representing the average preference partworths of the population; after the first question, Ψ is a matrix of eight columns and two rows, one row for the intercept and one row for the contribution of the first descriptor to preference partworths, etc.

4.4. Results

Table 5 reports the out-of-sample predictive accuracy of the three methods in the holdout task, which consisted of dividing 100 points among 4 alternatives. The first metric (*hit rate*) reports the frequency with which each method correctly predicts the participant’s top choice and indicates the method’s ability to identify the most likely preferred alternative. The second metric (*correlation*) reports Spearman’s rank correlation and is a proxy of the ability of each method to sort alternatives in order of customers’ preferences. Both metrics lead to the same conclusions.

As a benchmark, we report the in-sample predictive accuracy of classic conjoint analysis, where we used ratings from the main conjoint task without descriptors to predict holdout choices. This method uses the same respondents for conjoint data (estimation) and holdout data (prediction), whereas the other three methods use two different populations (training respondents for estimation and testing respondents for prediction) and, hence, should be regarded as a fair benchmark. The conjoint estimates achieved a hit rate of 0.575 (after 21 questions), an improvement of 130% compared to chance, and a rank correlation of 0.506.

The stepwise componential regression method, with a hit rate of 0.592 and a rank correlation of 0.502 achieved after only two questions, dominates the other methods, both in terms of number of questions asked and out-of-sample predictive accuracy (hit rate and correlation), and achieves a predictive accuracy not statistically different ($p < 0.01$) from the one the full conjoint study achieved after 21 questions (see Table 5 and Figure 3).

The closest contender to stepwise componential segmentation is Bayesian treed regression, with an average predictive accuracy of 0.528 at its end nodes, reached after 4.2 questions on average. Despite conservative parameterization and one-third of the training set dedicated to overfitting diagnosis, the treed regression approach suffered from overfitting problems, as shown in Figure 4: its maximum predictive accuracy was achieved after only two questions, with a hit rate of 0.580 and a rank correlation of 0.433, but it failed to stop the splitting process and its performance gradually deteriorated afterward.

These results can be compared to the out-of-sample accuracy of the stepwise componential segmentation method after an equal number of questions. Because

Table 4 Stepwise Componential Segmentation’s Preference Estimates After n Questions. Estimated Elements of Matrix Ψ

Step	Descriptors	Vector of preference partworths(*)							
		Intercept	1-day extended report	Sports news	\$4.00 coupon	Mix U.S./world news	World news only	Mix stocks/general news	General news only
Q1	Base	45.4	-7.2	-2.9	3.6	11.8	-8.1	13.5	8.2
	Base	40.4	-8.1	13.4	4.0	10.6	-9.1	10.9	5.8
	Descriptor 1 ^(a)	+7.5	+1.3	-24.7	-0.6	+1.9	+1.6	+4.0	+3.6
Q2	Base	33.5	7.5	14.0	4.3	10.0	-9.8	9.3	5.1
	Descriptor 1	+7.7	+0.9	-24.7	+0.6	+1.9	+1.5	+4.0	+3.6
	Descriptor 2 ^(b)	+10.7	-24.3	-0.9	-0.6	+1.0	+1.3	+2.5	+1.0
Q3	Base	38.3	4.8	15.4	3.8	10.8	-9.4	8.1	-3.7
	Descriptor 1	+8.9	+0.1	-24.2	-0.7	+2.6	+2.0	+3.6	+0.8
	Descriptor 2	+11.0	-24.5	-0.8	-0.5	+0.9	+1.5	+2.4	+0.5
	Descriptor 3 ^(c)	-7.5	+4.4	-2.3	+0.9	-1.6	-1.1	+2.1	+14.6

Notes. Elements in bold are significant at $p < 0.05$.

(*)The following attribute levels are set to 0 for identification purpose: 5-day forecast (weather report attribute), general news (university news), \$2 coupon (online coupon), U.S. news only (general news), stocks news only (business news).

^(a)Descriptor 1: Update as indicated if respondent answers “yes” to the question “General [university] news is typically more important to me than [university] sport news.”

^(b)Descriptor 2: Update as indicated if respondent answers between 4 (“Disagree”) and 6 (“Strongly disagree”) to the question “With regards to local weather reports, detailed summaries of today’s weather is typically more important to me than a less detailed five day forecast.”

^(c)Descriptor 3: Update as indicated if respondent answers “yes” to the question “General business news is typically more important to me than stock market news.”

preference partworths are computed on the entire data set, the method does not overfit the data. In contrast, Bayesian treed regression’s estimates are computed at the node level, i.e., on shrinking portions of the data set, leading to overfitting.

The cluster-classification method, with a hit rate of 0.482, a rank correlation of 0.386, and a much longer sequence of questions, fares far worse than the other two methods.

Table 5 Comparison of the Three Methods to Full-Profile Conjoint Analysis Averaged Over Tenfold Cross-Validation

	Full-profile conjoint	Cluster classification	Bayesian treed regression	Stepwise componential regression
Predictive accuracy, hit rate (%)	57.5	48.2	52.8	59.2
Predictive accuracy, correlation (%)	50.6	38.6	43.3	50.2
Average number of questions	21	4.0	4.2	2
Maximum number of questions	21	11	6	2
Incremental gain in predictive accuracy, per question (*) (%)	1.5	5.8	6.6	17.1

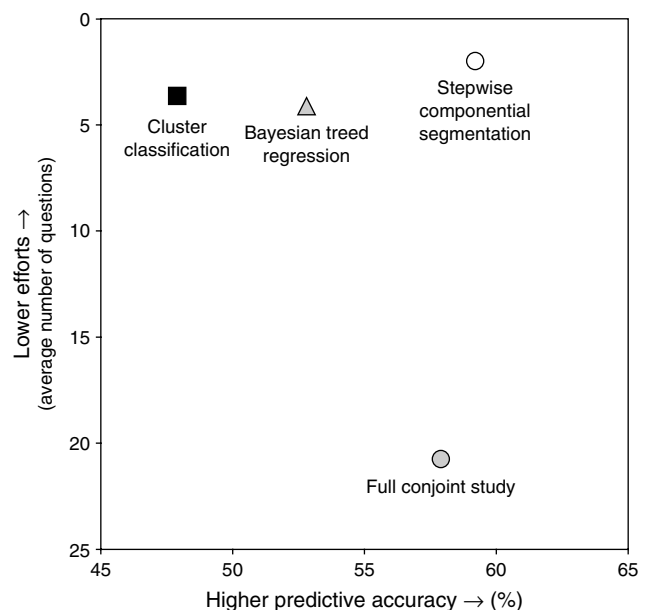
Notes. The stepwise componential regression method’s predictive accuracy is similar ($p < 0.01$) after two easy-to-answer questions to the predictive accuracy of classic, full-profile conjoint analysis after 21 profile rating questions. Cluster-classification and Bayesian treed regression methods are dominated, achieving lower predictive accuracy while requiring a higher number of questions than stepwise componential segmentation.

(*) = (predictive accuracy – 25%)/number of questions. 25% is what is achieved by chance.

4.5. The Question Selection Mechanism

For the componential segmentation method, the one that performed best, a short sequence of questions achieves the same predictive accuracy as a full classic

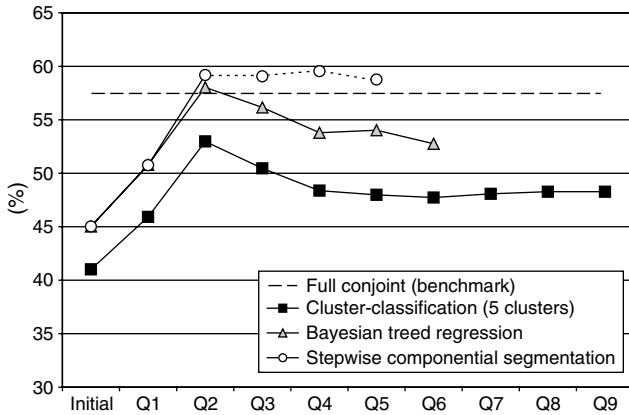
Figure 3 Methods Plotted on Out-of-Sample Predictive Accuracy (i.e., Hit Rate) and Effort Required (i.e., Number of Questions)



Notes. The stepwise componential segmentation, in the upper right corner, dominates the cluster classification and Bayesian treed regression methods on both dimensions and achieves a predictive accuracy not statistically different from that achieved by the conjoint study ($p < 0.01$), with far fewer questions.

Copyright: INFORMS holds copyright to this Article in Advance version, which is made available to institutional subscribers. The file may not be posted on any other website, including the author’s site. Please send any questions regarding this policy to permissions@informs.org.

Figure 4 Out-of-Sample Predictive Accuracy (Hit Rate) of the Three Competing Methods After n Questions



Notes. Both stepwise componential segmentation and Bayesian treed regression achieve excellent predictive accuracy but the latter eventually suffers from overfitting. Notice that the stepwise componential segmentation algorithm suggests stopping after two questions—Predictive accuracy afterward is only reported for comparison purposes.

conjoint study. The method selects those by testing all possible questions (i.e., splits) and retaining those leading to highest performance improvement. We analyze the underlying statistical reasons that make some questions more valuable than others in predicting preference partworths using a hierarchical Bayesian model.

We extend the original *componential segmentation* modeling framework by using an analogous hierarchical Bayesian model. The probability model is the standard regression model with a normal error, or

$$y_{ij} =_d N(\beta_i \cdot P_{ij}, \sigma^2), \tag{3}$$

where $=_d$ means equal in distribution and N is the normal distribution. We assume standard hierarchical priors, such that

$$\sigma^2 =_d IG(Shape, Scale), \tag{4}$$

where IG is the inverse gamma density and *Shape* and *Scale* are chosen to reflect vague prior information. In addition, we assume that

$$\beta_i =_d N(\psi \cdot D_i, \Lambda), \tag{5}$$

where each element of the aggregator matrix is a priori independent and normally distributed, and

$$\psi_{kq} =_d N(\bar{\psi}_{kq}, \tau), \tag{6}$$

with hyper priors which reflect vague prior information; a priori the heterogeneity matrix

$$\Lambda =_d IWishart(P, n) \tag{7}$$

is assumed to follow an inverse Wishart distribution again with vague prior information. Inference about

Ψ and other parameters is obtained using standard Markov chain Monte Carlo (MCMC) methods; see Gilks et al. (1996) for a discussion of MCMC methods.

This specification results in a shrinkage model, where each individual partworth estimate is shrunk toward an individual specific mean which is the result of a linear combination of the aggregator matrix and the descriptor variables, or $\psi \cdot D_i$. If the descriptor variables are not individual specific, e.g., $D_i = 1$, then each individual will have the same mean partworth value and the model reduces to the standard hierarchical shrinkage model. In the context of prediction and understanding consumer heterogeneity, the following observations are important.

First, the proposed hierarchical model fits into our overall framework in Figure 1 by allowing us to easily use descriptor variables to estimate partworths (and preferences) for the new individuals who are participating in the Web dialog. An estimate of the posterior mean of the attractor matrix $\hat{\psi}$ that has been formed based on analysis of an ex ante set of conjoint data can be used to form an estimate of a new individual's partworth, given that we know the descriptor variables D_i for that individual by $\hat{\beta}_i = \hat{\psi} \cdot D_i$. As with the classic stepwise componential segmentation approach, this individual-level estimate can be used to form *predicted* preference scores \tilde{y} without asking the new individual to evaluate a product alternative.

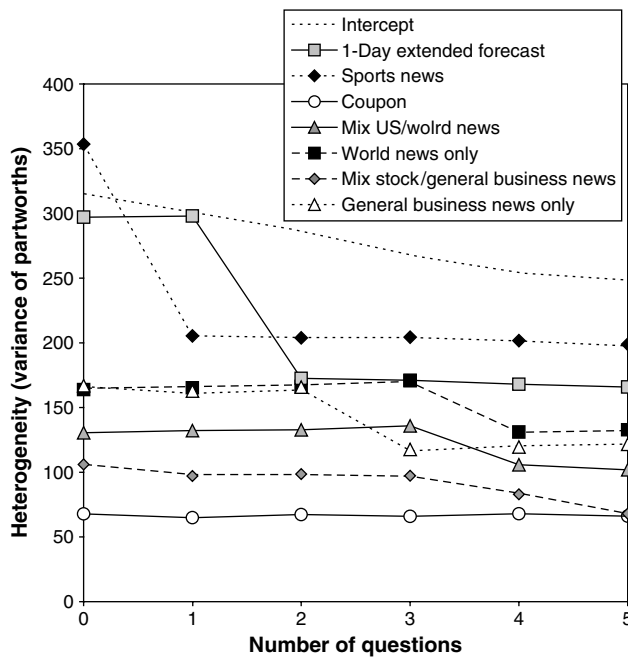
A second observation concerns the size of the diagonal elements of the heterogeneity matrix Λ ; the extent to which the descriptor variables explain individual partworth preferences is reflected in the amount of observed heterogeneity or in the size of the diagonal elements of Λ . These diagonal elements reflect the variance of the error between the individual mean partworth, given by $\psi \cdot D_i$, and the actual partworth; as these errors decrease or as the individual mean and actual partworths become closer to each other, this variance term goes to zero.

Hence, using this Bayesian model, the strength of the relationship between descriptor variables and individual partworths can be seen in terms of the change in the heterogeneity of a partworth (summarized by the diagonal elements of Λ), when a descriptor variable is included in the stepwise componential segmentation analysis. If the newly introduced descriptor variable helps predict a partworth for a set of individuals better, then we would expect the heterogeneity to decrease.

As demonstrated in Figure 5, the heterogeneity for the partworths change as new descriptor variables are added (we use the first cross-validation sample for illustration throughout). It appears that the inclusion of a descriptor question mostly impacts a single partworth, with the exception of the fourth descriptor

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to institutional subscribers. The file may not be posted on any other website, including the author's site. Please send any questions regarding this policy to permissions@informs.org.

Figure 5 Heterogeneity (Diagonal of Λ for the Stepwise Componential Segmentation Analysis) for Each Partworth as a Function of the Number of Descriptor Questions That Were Included in the Analysis



Notes. The questions selected by the *stepwise componential segmentation* method reduce the heterogeneity of the most heterogeneous partworths.

question. The first descriptor question (“General [university] news is typically more important to me than [university] sport news”) reduces the heterogeneity of the news partworth by almost 50%. The second descriptor question (“With regard to local weather reports, detailed summaries of today’s weather is typically more important to me than a less detailed five-day forecast”) reduces the heterogeneity of the partworth for one-day extended forecast. The third descriptor question (“General business news is typically more important to me than stock market news”) reduces the heterogeneity of the general business news. These impacts make intuitive sense.

Note that the first question selected helps reduce the variance of the most heterogeneous partworth the second question is related to the second most heterogeneous partworth and so on. Hence, the question selection method appears to be an *efficient heterogeneity reduction* mechanism which can be achieved best when (a) partworth heterogeneity for an attribute is large in the population, and (b) preferences for this attribute are correlated to available descriptor questions.

While heterogeneities continue to decrease as descriptor questions are added, the predictive power of the method stops improving after including the first and second descriptor, for two reasons. First, the reductions in heterogeneity are much larger for

the first two descriptor questions, suggesting that they have a larger impact on predicting these individual partworths. Second, the remaining partworths exhibit some attribute correlation with these initial attributes, reducing the remaining unexplained heterogeneity.

4.6. Sensitivity Analysis

Stepwise componential segmentation required only two questions to achieve the same predictive accuracy as a full, 21-profile rating conjoint study. We now explore the extent to which such results are sensitive to these two specific questions. In other words, if we do not include those items in the initial pool of questions used to develop the decision aid, how much does the predictive performance decrease? This question is critical, because it indicates the extent to which the method’s ability to recover consumers’ preferences is sensitive to the initial pool of questions.

Table 6 reports the first four most informative questions from the entire set retained by the *stepwise componential segmentation* method. The elements in the table indicate the number of times each question appeared for each cross-validation sample and its position. Across all cross-validations, the same first four questions were retained in approximately the same order.

The four most informative questions are the four self-explicated items introduced in the questionnaire, which tap directly into the respondents stated preferences. Although participants’ answers point directly to their most preferred features, the *ex ante* conjoint

Table 6 Questions Selected by the Stepwise Componential Segmentation Algorithm to Elicit Customers’ Preferences, Across All 10 Cross-Validations

	First question	Second question	Third question	Fourth question
Question N°48: “With regards to local weather reports, detailed summaries of today’s weather is typically more important to me than a less detailed five-day forecast.”	4	6		
Question N°56: “General [university] news is typically more important to me than [university] sport news.”	6	4		
Question N°71: “National news is typically more important to me than world news.”			1	9
Question N°79: “General business news is typically more important to me than stock market news.”			9	1

Notes. Question 48 was identified four times as being the most informative question to ask, and six times as being the second most informative question to ask.

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to institutional subscribers. The file may not be posted on any other website, including the author’s site. Please send any questions regarding this policy to permissions@informs.org.

analysis was necessary to (a) identify in which order these questions had to be asked to extract preference information as efficiently as possible, and (b) identify the most likely level of respondents' preferences for attributes in order to predict respondents' choices when they had to choose among different suboptimal profiles. Note that the first two most informative questions pertain to two nondominant attributes, explaining 15% and 17% of the variance each (see Table 1). The method identified these two attributes to be those where respondents could be discriminated most efficiently. For the three remaining attributes (face value of an online coupon, general news, and business news), the method found that offering the \$4 online coupon and a mix of both general news and business news would satisfy most respondents. For instance, although 34% of respondents preferred general business news only and 8% preferred stocks news only (see Table 1), to offer all respondents a mix of general business/stocks news would usually incur a negligible loss in preferences. Hence, incremental efforts needed to identify those with more marked preferences for one type of business news were not justified by the potential small gains in utility, and the method did not incorporate questions related to this dimension.

In some applications, either due to lack of domain knowledge or to lack of data availability, self-explicated questions may not be available. Figure 6 reports the predictive accuracy of the stepwise componential segmentation algorithm, either with the full list of 99 available questions (already reported in Figure 4) or with a restricted list of 95 questions, of which self-explicated items had been removed. The method stops after four questions, with a hit rate of 0.496 and an average predictive accuracy gain per question of 6.2% (compared to 17.1% for the same method with

self-stated preference questions available and 1.5% for the full conjoint analysis). The restricted pool of questions has a significantly lower information value ($p < 0.01$) and predictive accuracy increases at a much slower rate, indicating the importance of selecting good questions.

5. Understanding the Results

In this research, we used the in-sample predictive accuracy of a full conjoint analysis as a benchmark to compare the out-of-sample performance of three different implementations of conjoint-based decision aids. This benchmark tells how well customers' preferences can be elicited, and how valuable preference models used in conjoint analysis can be with respect to measuring and predicting which products customers will prefer.

The stepwise componential segmentation methodology required only two questions where full conjoint required 21, and yet achieved similar predictive accuracy on both hit rate and correlation performance metrics. Although this outstanding performance may seem surprising, it can be explained by considering the nature of the information that is captured by the descriptor questions.

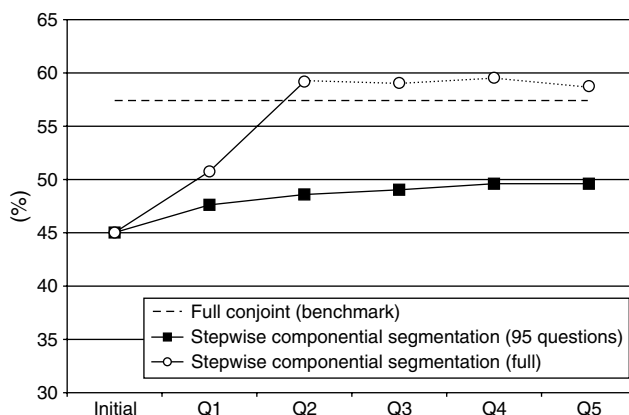
Conjoint-based decision aids presented in this research did not actually generate any knowledge about respondents' preferences. This knowledge base was already available from a sample of the population, generated during an ex ante conjoint study. The role of the decision aid was simply to retrieve the preferences as efficiently (from a customer's standpoint) as possible. This knowledge retrieval process can be accelerated by exploiting the inefficiencies that occur in most conjoint studies and that are probably unavoidable in most practical settings because of:

Unimportant Attributes. If certain attributes are irrelevant or unimportant to explain customers' choices or for a subset of those customers, the decision aid may not need to explore the underlying reasons that drive preferences for such attributes, avoiding substantial unnecessary data collection.

Dominated Levels. Certain attribute levels might be dominated, that is, least preferred by most customers, hence reducing questionnaire length substantially. For instance, in their decision to buy a laptop, most customers might be willing to pay extra to have embedded wireless internet capabilities, and asking questions about customers' needs in terms of mobility might be of no real added value.

Attribute Correlations. While in theory attribute preferences in a conjoint study should not be correlated, this criterion is rarely achieved in practice. Consequently, well-designed questions could potentially inform the decision aid of preferences for several

Figure 6 Out-of-Sample Predictive Accuracy (Hit Rate) of the Stepwise Componential Segmentation, with a Full or Restricted Version of the Questionnaire



Notes. The method is overly sensitive to the initial pool of questions available to draw inference about respondents' preferences.

attributes at once and, in doing so, increase questionnaire efficiency. For instance, the question “do you intend to use your computer to play 3D video games?” may be valuable to simultaneously predict preferences for screen size, video card quality, processor speed, speaker system, and so on.

The decision aid can include straightforward, self-explicated questions (e.g., “do you need your laptop to be wireless-Internet enabled?”) or even product specification questions. If no particular product category expertise is required to answer such questions or if customers have this expertise, such questions may represent the most efficient way to elicit preferences and make sound recommendations. Such questions are also often used in current Web-based recommendation systems. However, the system we propose uses self-stated preferences to weight rather than to screen alternatives, an approach that offers the following useful properties.

First, the system will discard self-reported preference questions about characteristics that either do not matter (unimportant attributes) or for which the most likely answer is already known (dominated levels), while retaining those questions with the highest potential impact on predictive accuracy (important attributes with high heterogeneity of preferences in the population), hence increasing questionnaire efficiency.

Second, even straightforward self-reported preferences for one attribute might help predict preferences for other attributes. For instance, a customer who would state that she “very much” needs a wireless-Internet enabled laptop might also signal that she needs a highly mobile computer system. The decision aid might therefore infer, based on the results of the ex ante conjoint study, that she will give more preference weights to smaller, lighter laptops with better-than-average battery autonomy. The ability to identify

questions that predict preferences for more than one attribute at a time could be an important benefit of such methods and occurs due to correlations in preferences across attributes. The number of such significant correlations is likely to increase as the number of attributes increases, suggesting that the relationship between the number of attributes and the number of questions required to elicit preferences might be concave.

Third, conjoint-based decision aids assign weights based on the analysis of past customers’ preferences who answered similarly. They use those weights to rank alternatives and offer compromises beyond self-reported preferences. For instance, a customer who does not know the laptop market may specify that he is looking for a computer that weighs less than 5 pounds and costs less than \$500—an unrealistic combination at the time of this writing. A classic specification-based recommender system would be unable to recommend any product, and would request the customer to choose which constraints should be relaxed and with which magnitude until a nonempty set of recommendations is found; a conjoint-based decision aid would exploit the results of the ex ante conjoint study to predict which compromise is least likely to affect customers’ preferences and make recommendations accordingly.

The power of using descriptor questions to estimate an individual partworth is related to the correlation between descriptor questions and partworth utilities. The relative advantages and disadvantages of each type of question (from most specific to most general) are summarized in Figure 7. Self-stated preferences (e.g., product specification) are likely to predict preferences for a specific attribute with great precision while more general descriptors (e.g., demographics), which influence various choice dimensions, will be able to predict preferences for a wider range of

Figure 7 Specific, Self-Reported Preferences Will Have a High Predictive Accuracy on One or a Few Attributes at a Time, While More General Questions (e.g., Demographics) Will Contribute to Predict Preferences for a Broader Range of Product Characteristics but with Less Precision

Type of question	Example 1 Laptop, wireless internet	Example 2 Car security	Predictive accuracy
Product specification	Which option do you prefer? [Wi-Fi modem 110.11 b/c/g + Bluetooth; Wi-Fi modem 110.11 b/c only; etc.]	Which option do you prefer? [2 frontal + 4 lateral airbags; 2 frontal airbags only; etc.]	
Self-reported preferences	What wireless internet capabilities do you expect from your laptop? [None; ... Latest technologies]	Is the number of airbags a key element in your decision to buy a car? [Not at all; ... Very much]	
Product usage	Will you often work on the move with your computer?	How many miles per year will you most likely drive the car?	
Demographics	What is your profession?	Do you have children?	
			Medium-low, broad

Notes. For instance, the question “Do you have children?” might be a good indicator for airbag preferences but probably less accurate than other, more specific questions. However, it might also help predict budget constraints, preferences for family cars, etc. Conjoint-based decision aids will select the most appropriate questions within a specific context.

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to institutional subscribers. The file may not be posted on any other website, including the author's site. Please send any questions regarding this policy to permissions@informs.org.

attributes but probably less precisely. Depending on the context (product category, customers' expertise, existing attribute correlations, etc.), the conjoint-based decision aid will identify certain types of questions as being more efficient and will use them more often to design optimal questionnaires and recommendations.

6. Discussion and Conclusions

In their purchase decisions, customers try to improve decision quality while limiting search efforts. On the Internet, merchant Web sites have seized the opportunity offered by the electronic environment to offer decision aids and recommender systems to their customers. These systems help customers make more efficient use of their search time. In this context, we have proposed a framework by which companies can develop recommendation agents that are capable of providing high quality advice to customers with minimum input. Specifically, we explored how the richness of preference models used in traditional conjoint analysis techniques could be leveraged to design online decision aids without requiring the extensive and detailed inputs usually necessary for these kinds of models. We tested alternative implementations of the approach and have shown that the stepwise componential regression method achieved the same predictive accuracy as a full conjoint analysis while offering great efficiency gain (predictive accuracy increased by 17.1% per question, compared to 1.5% for conjoint analysis).

Our results relied on a single data set in a specific context; future work should assess the performance in predictive accuracy and efficiency that one can expect to achieve in other contexts. Clearly, to be applicable at all, the approach requires a multiattribute choice context amenable to conjoint analysis, and will likely be most appropriate when customers' choices are mainly driven by objective and identifiable needs and constraints (e.g., computers, video cameras, cars, financial products, etc.). If alternatives can not be easily described by a finite set of known attributes (e.g., books, movies, music), other methods such as collaborative filtering might remain the most effective way to make recommendations in those contexts. The approach is also likely to be less applicable where the stability of preferences is low, such as with fashion goods.

It would be valuable to assess the usability of our method when the number of attributes and attribute levels increase. On the one hand, the ex ante conjoint study would increase in complexity and cost and the resulting questionnaire would also increase in length. On the other hand, as the number of attribute and attribute levels increases, we can expect: (a) an increasing proportion of attributes of lesser

importance, (b) an increase in the number of dominated levels, and (c) higher correlations among attributes, three factors that underlie the effectiveness of our proposed method. Consequently, the relationship between product complexity and questionnaire length might be concave, enabling our approach to be effective in such situations.

The pool of questions used is likely to be critical to the performance of the method. Wedel and Kamakura (2000) suggest that the success of stepwise componential regression will mostly depend on the existing correlations between customers' descriptors and individual preference partworths. If these correlations are weak, recommendations will be unsatisfactory. We have shown that dropping four key questions decreased the gains in average incremental predictive accuracy threefold (from 17.1% to 6.2%). *Stepwise componential segmentation*, in particular, works as a heterogeneity-reduction mechanism, identifying questions that diminish heterogeneity around cluster centroids. If such questions can not be identified (whether because preferences are loosely correlated with descriptors or because relevant questions have not been included in the set of questions asked), heterogeneity around the centroid will remain high and recommendation quality will not significantly improve.

In our study, self-explicated questions were the most informative and the method easily identified them. This ease of identification might not always be the case. Self-explicated questions might be too complex or might lead to poor performance, depending on the complexity of the product category or the limited product knowledge or experience of the customer. When self-explicated questions are not appropriate, other methodologies should be explored to identify the best questions to include in the questionnaire. Although we used a pilot study and experts, other methods such as ethnography and observational research, group/brainstorming sessions and focus groups, laddering (i.e., means-end chaining), and scripting and cognitive process interviews should be studied to see when and why they generate effective questions.

Both *stepwise componential segmentation* and *cluster classification* methods are locally optimal methods, i.e., they select the next most informative question locally. Much as in stepwise regression, there might be other sequences of questions that are globally superior to those selected by these locally-optimal approaches. *Bayesian treed regression*, using global search, does not suffer from this limitation but suffers from overfitting issues. Amending the proposed method to allow for global search represents an interesting methodological challenge.

Diehl et al. (2003) and Lynch and Ariely (2000) have shown that lowering quality search costs can increase consumers' price sensitivity. Consequently, the implementation of a better and easier-to-use recommendation agent might not lead to an increase in the firm's profits. In our context, it might be of interest to incorporate a company's cost and profit structure in our proposed recommender system so that the decision aid could explicitly weight customers' preferences (i.e., most preferred products) and firms' interests (i.e., higher-margin products) in making recommendations. Such a profit-maximizing recommender agent would require (a) transforming preference scores into choice likelihood—a rather straightforward addition given the state of choice modeling—and (b) the computation of a profit function for the firm based on products' characteristics. This approach might enable a merchant Web site to make informed, profit-driven product recommendations to their online visitors.

To enable wide use of methods such as those suggested here, one should consider approaches that do not require merchant Web sites to conduct regular ex ante conjoint studies to feed and update their recommender systems. Other ways to generate a knowledge base should be explored such as hybrid conjoint methods, rating mechanisms, or unobtrusive online data collection (where those products browsed and bought by a given customer are used to build simulated choice-based conjoint data points). For instance, Weng and Liu (2004) use past purchase data to infer customers' preferences for certain features (although they do not link these preferences to consumers' profiles and, hence, can not make relevant recommendations to first-time visitors).

Aksoy et al. (2006) have shown that consumers made much better choices when they used electronic agents that "thought like people," both in terms of search strategy and attribute weights. Kamis and Stohr (2006) also found that perceived ease of use of recommender systems influenced decision confidence, perceived usefulness, decision quality, and, ultimately, acceptance of the electronic agent. Our conjoint-based decision aids meet these requirements.

Conjoint-based decision aids offer opportunities both for future research and promising real-life applications. Given the increasing presence of online merchant catalogs and the rise of online product customization, this type of recommender system might very well find its place among the suite of alternatives available to help customers make better and more efficient purchase decisions.

Acknowledgments

The authors would like to thank the Institute for the Study of Business Markets at The Pennsylvania State University

for its financial support. This paper is based on the first author's Ph.D. thesis at Penn State University.

References

- Aksoy, L., P. N. Bloom, N. H. Lurie, B. Cooil. 2006. Should recommendation agents think like people? *J. Service Res.* 8(4) 297–315.
- Ansari, A., S. Essegai, R. Kohli. 2000. Internet recommendation systems. *J. Marketing Res.* 37(August) 363–375.
- Balasubramanian, S. K., W. A. Kamakura. 1989. Measuring consumer attitudes toward the marketplace with tailored interviews. *J. Marketing Res.* 26(3) 311–326.
- Banister, E. N. 2003. Using the Internet for customer behaviour research: A guide to the challenges and opportunities of the Internet as a survey medium. *J. Customer Behav.* 2(1) 105–123.
- Bergmann, R., S. Schmitt, A. Stahl. 2002. Intelligent customer support for product selection with case-based reasoning. P. S. S. Javier Segovia, M. Niedzwiedzinski, eds. *E-Commerce and Intelligent Methods*. Physica-Verlag, Heidelberg, New York, 322–341.
- Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth, New York.
- Chipman, H., E. George, R. McCulloch. 2002. Bayesian treed models. *Machine Learn.* 48 299–320.
- Choi, S. H., S. Kang, Y. J. Jeon. 2006. Personalized recommendation system based on product specification. *Expert Systems Appl.* 31 607–616.
- Chung, G. 2004. Developing a flexible spoken dialog system using simulation. *Proc. 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain (July) 955–964.
- DeLong, C., P. Desikan, J. Srivastava. 2005. USER (User Sensitive Expert Recommendation): What non-experts NEED to know. *Proc. WebKDD2005 (Workshop on Knowledge Discovery in the Web)*, Chicago, IL.
- DeSarbo, W. S., R. L. Oliver, A. Rangaswamy. 1989. A simulated annealing methodology for clusterwise linear regression. *Psychometrika* 54(December) 707–736.
- DeSarbo, W. S., M. Wedel, M. Vriens. 1992. Latent class metric conjoint analysis. *Marketing Lett.* 3(July) 273–289.
- DeSarbo, W. S., A. M. Degeratu, M. J. Ahearne, K. M. Saxton. 2002. Disaggregate market share response models. *Internat. J. Res. Marketing* 19 253–266.
- Diehl, K., L. J. Kornish, J. G. Lynch. 2003. Smart agents: When lower search costs for quality information increase price sensitivity. *J. Consumer Res.* 30(1) 56–71.
- Gilks, W. R., S. Richardson, D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Green, P. E. 1984. Hybrid models for conjoint analysis: An expository review. *J. Marketing Res.* 21(May) 155–169.
- Green, P. E., W. S. DeSarbo. 1979. Componential segmentation in the analysis of consumer trade-offs. *J. Marketing* 43(4) 83–91.
- Green, P. E., A. M. Krieger. 1991. Segmenting markets with conjoint analysis. *J. Marketing* 55(October) 20–31.
- Green, P. E., A. M. Krieger, C. M. Schaffer. 1993. An empirical test of optimal respondent weighting in conjoint analysis. *J. Acad. Marketing Sci.* 21(Fall) 345–351.
- Grenci, R. T., P. A. Todd. 2002. Solutions-driven marketing. *Comm. ACM* 45(3) 65–71.
- Gupta, S., P. K. Chintagunta. 1994. On using demographic variables to determine segment membership in logit mixture models. *J. Marketing Res.* 31(1) 128–136.
- Hagen, P., D. E. Weisman, H. Manning, R. K. Souza. 1999. *Guided Search for E-Commerce*. Forrester Research, Inc., Cambridge, MA.
- Hagerty, M. R. 1985. Improving the predictive power of conjoint analysis: The use of factor analysis and cluster analysis. *J. Marketing Res.* 22(May) 168–184.

- Häubl, G., V. Trifts. 2000. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Sci.* **19**(1) 4–21.
- He, C., Y. Chen. 2006. Managing e-marketplace: A strategic analysis of nonprice advertising. *Marketing Sci.* **25**(2) 175–187.
- Herlocker, J. L., J. A. Konstan, J. Riedl. 2000. Explaining collaborative filtering recommendations. *Proc. 2000 ACM Conf. Comput. Supported Cooperative Work*, Philadelphia, PA, 241–250.
- Hoch, S. J., D. A. Schkade. 1996. A psychological approach to decision support systems. *Management Sci.* **42**(1) 51–64.
- Huberty, C. J. 1989. Problems with stepwise methods: Better alternatives. B. Thompson, ed. *Advances in Social Science Methodology*, Vol. 1. JAI Press, Greenwich, CT.
- Huberty, C. J. 1994. *Applied Discriminant Analysis*. Wiley and Sons, New York.
- Huffman, C., B. Kahn. 1998. Variety for sale: Mass customization or mass confusion? *J. Retailing* **74**(4) 491–513.
- Johnston, M., S. Bangalore, G. Vasireddy. 2001. MATCH: Multimodal access to city help. *Automatic Speech Recognition and Understanding Workshop*. Trento, Italy.
- Kamakura, W. A. 1988. A least square procedure for benefit segmentation with conjoint experiments. *J. Marketing Res.* **25**(May) 157–167.
- Kamakura, W. A., M. Wedel. 1995. Life-style segmentation with tailored interviewing. *J. Marketing Res.* **32**(August) 308–317.
- Kamakura, W. A., M. Wedel, J. Agrawal. 1994. Concomitant variable latent class models for conjoint analysis. *Internat. J. Res. Marketing* **11**(5) 451–464.
- Kamis, A. A., E. A. Stohr. 2006. Parametric search engines: What makes them effective when shopping online for differentiated products? *Inform. Management* **43** 904–918.
- Kaufman, L., P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, Hoboken, NJ.
- Konstan, J. A., B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon. 1997. GroupLens: Applying collaborative filtering to Usenet news. *Comm. ACM* **40**(3) 77–87.
- Kuflik, T., B. Shapira, P. Shoval. 2003. Stereotype-based versus personal-based filtering rules in information filtering systems. *J. Amer. Soc. Inform. Sci. Tech.* **54**(3) 243–250.
- Lynch, J. G., D. Ariely. 2000. Wine online: Search costs affect competition on price, quality, and distribution. *Marketing Sci.* **19**(1) 83–103.
- Netflix. 2006. Netflix prize. <http://www.netflixprize.com>.
- Ogawa, K. 1987. An approach to simultaneous estimation and segmentation in conjoint analysis. *Marketing Sci.* **6**(1) 66–81.
- O’Keefe, R. M., T. McEachern. 1998. Web-based customer decision support systems. *Comm. ACM* **41**(3) 71–78.
- Popescul, A., L. H. Ungar, D. M. Pennock, S. Lawrence. 2001. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. M. Kaufmann, ed. *Proc. 17th Conf. Uncertainty Artificial Intelligence*, San Francisco, CA.
- Randall, T., C. Terwiesch, K. T. Ulrich. 2005. Principles for user design of customized products. *California Management Rev.* **47**(4) 68–85.
- Rencher, A. C. 1995. *Methods of Multivariate Analysis*. John Wiley & Sons, New York.
- Resnick, P., H. R. Varian. 1997. Recommender systems. *Comm. ACM* **40**(3) 56–58.
- Sawtooth Software, Inc. 2002. *ACA 5.0 Technical Paper*, Sawtooth Software, Inc., Sequim, WA.
- Schafer, J. B., J. A. Konstan, J. Riedl. 2001. E-commerce recommendation applications. *Data Mining Knowledge Discovery* **5**(1/2).
- Shardanand, U., P. Maes. 1995. Social information filtering: Algorithms for automating “Word of Mouth.” *Proc. ACM CHI’95 Conf. Human Factors Comput. Systems*, Denver, CO.
- Shugan, S. M. 1980. The cost of thinking. *J. Consumer Res.* **7**(September) 99–111.
- Simon, H. H. 1957. *Models of Man*. John Wiley & Sons, Inc., New York.
- Singh, J., R. Howell, G. K. Rhoads. 1990. Adaptive designs for Likert-type data: An approach for implementing marketing surveys. *J. Marketing Res.* **27**(August) 304–321.
- The Washington Post*. 2006. Wal-Mart web site makes racial connections: DVD shoppers get offensive referrals. (January 6).
- Toubia, O., D. I. Simester, J. R. Hauser, E. Dahan. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Sci.* **22**(3) 273–303.
- Tran, T. 2006. Designing recommender systems for ecommerce: An integration approach. *Proc. 8th Internat. Conf. on Electronic Commerce (ICEC)*, New Brunswick, Fredericton, Canada, 512–518.
- Vriens, M., M. Wedel, T. Wilms. 1996. Metric conjoint segmentation methods: A Monte Carlo comparison. *J. Marketing Res.* **32** 73–85.
- Wedel, M., W. A. Kamakura. 2000. *Market Segmentation, Conceptual and Methodological Foundations*, 2nd ed. Kluwer Academic Publishers, Boston, MA.
- Wedel, M., C. Kistmaker. 1989. Consumer benefit segmentation using clusterwise linear regression. *Internat. J. Res. Marketing* **6** 45–49.
- Wedel, M., J.-B. E. M. Steenkamp. 1989. A fuzzy clusterwise regression approach to benefit segmentation. *Internat. J. Res. Marketing* **6**(March) 45–59.
- Wedel, M., J.-B. E. M. Steenkamp. 1991. A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation. *J. Marketing Res.* **28**(November) 385–396.
- Weng, S.-S., M.-J. Liu. 2004. Feature-based recommendations for one-to-one marketing. *Expert Systems Appl.* **26** 493–508.
- Wittink, D. R., M. Vriens, V. Burhenne. 1994. Commercial use of conjoint analysis in Europe: Results and critical reflections. *Internat. J. Res. Marketing* **53** 41–52.
- Wright, P. 1975. Consumer choice strategies: Simplifying vs. optimizing. *J. Marketing Res.* **12**(February) 60–67.