Active Learning for Consideration Heuristics

by

Daria Dzyabura
and
John R. Hauser

January 2009

Daria Dzyabura is a doctoral student at the MIT Sloan School of Management, Massachusetts Institute of Technology, E40-170, One Amherst Street, Cambridge, MA 02142, (617) 253-2268, dariasil@mit.edu.

John R. Hauser is the Kirin Professor of Marketing, MIT Sloan School of Management, Massachusetts Institute of Technology, E40-179, 1 Amherst Street, Cambridge, MA 02142, (617) 252-2929, hauser@mit.edu.

This research was supported by the MIT Sloan School of Management, the Center for Digital Business at MIT (ebusiness.mit.edu), and General Motors, Inc. We would like to thank Rene Befurt (Analysis Group), Sigal Cordeiro (GM), Theodoros Evgeniou (INSEAD), Patricia Hawkins (GM), Phillip Keenan (GM), Andy Norton (GM), Ele Ocholi (MIT), Kevin Wang (MIT), Erin MacDonald (MIT), Daniel Roesch (GM), Joyce Salisbury (GM), Matt Selove (MIT), Olivier Toubia (Columbia), Paul Tsier (MIT), Catherine Tucker (MIT), Glen Urban (MIT), and Juan-Juan Zhang (MIT) for their insights, inspiration, and help on this project.

Active Learning for Consideration Heuristics

Abstract

We propose and test an active-learning algorithm to estimate heuristic decision rules that consumers use to form consideration sets for automobile purchases. The complexity of the product (large number of features and levels) and the non-linearity in models of heuristic decisions lead to computational and data-collection challenges. For example, optimal orthogonal designs would require consumers to evaluate thousands of vehicles. We propose a querying strategy that chooses each question (consider vs. not consider) to minimize the expected posterior entropy. We use a variational-Bayes belief-propagation algorithm to perform estimation after each question so we rapidly find the optimal next question. Priors are consumer-specific and based on a configurator and market-level data.

Synthetic data experiments suggest that our proposed algorithm recovers information much more efficiently than standard approaches. We test the algorithm empirically in a web-based survey conducted by an American automotive manufacturer to study vehicle consideration (872 respondents, 53 aspects, 29 questions), and demonstrate that it significantly outperforms standard approaches on a variety of performance measures. We close by suggesting how the active-learning framework might be applied to other marketing science problems.

Keywords: *Active learning, automotive applications, belief-propagation, conjoint analysis, conjunctive models, consideration sets, consumer heuristics, decision heuristics, lexicographic models, variational Bayes estimation.*

1. Introduction and Motivation

Many methods have been proposed and tested to estimate heuristic decision rules from extant data, but as applications expand to complex products with many alternatives, many product features, and many respondents, applications face potentially onerous data collection and computational demands. To simplify data collection and computation we design and apply an active-learning framework to infer a consumer's decision heuristic. The system re-estimates the model after observing a consumer's answer using a variational-Bayes belief-propagation algorithm. It selects the next product profile for the consumer to evaluate such that we get the most information from each response.

We test the methodology on a web-based survey conducted in order to study consideration-set formation among US auto buyers. We focus on heuristic rules for consideration decisions for several reasons. (1) Behavioral science suggests that decision heuristics are common, especially when consumers evaluate many products or features (e.g., Bettman, Luce and Payne 1998; Gigerenzer and Goldstein 1996; Payne, Johnson and Bettman 1988, 1993). (2) Managerial interest is high among US auto makers because they perceive that the collapse of the US auto industry was due, in part, to the fact that today's US consumers are much less likely to consider GM, Ford, and Chrysler brands. (3) Without active queries it is extremely difficult and expensive to identify decision heuristics from observed consideration decisions. The $21x9x7x5x3^2x2$ design in our application requires 13,320 profiles for a (D-efficient) orthogonal design. And (4) decision heuristics are non-linear and defined on a discrete space. There are many published methods (reviewed below) to estimate decision heuristics from passive data when the design is moderate, but the methods become computationally impractical for large designs. For example, the cardinality of the space of conjunctive decision rules for the design in our application is 9.0×10^{15}. More complex decision rules would imply substantially higher cardinality. The space can be made continuous with probabilities, but, even so, the challenge is quite different from that posed by additive-partworth conjoint-analysis models.

Synthetic-data experiments on a smaller problem (4^4), for which orthogonal designs are feasible, suggest that the proposed algorithm recovers information more efficiently than standard approaches. To retrieve the same amount of information as a 32-profile orthogonal design, the proposed method requires just 9 questions. Randomly-chosen profiles require 40 questions and profiles chosen randomly but proportional to market share require 38 questions. In the empirical

study, we compare the active-learning-based questions to market-share-based questions because they are expected to outperform purely random questions and because orthogonal designs are not feasible for the empirical problem.

We begin with a brief review of active learning, decision heuristics for consideration sets, and existing methods to estimate such heuristics. We then describe a consumer-response model of the data, an active-learning framework, how we update beliefs after each query, and how we select queries. We compare active learning to commonly-used querying strategies with synthetic data experiments. We then illustrate and test the active-learning method in a study with 872 real consumers evaluating vehicle profiles (cars, trucks, SUVs, and minivans).

2. Review of Related Literatures

Active Learning

Active learning is a subfield of machine learning that focuses on optimally choosing data from which to train an algorithm (for a review, see Settles 2009). By choosing the most informative data points, or queries, an active-learning algorithm achieves greater accuracy with less training data. Typically, the algorithm is endowed with feasible prior information, chooses one or more carefully selected queries, learns from the query results, and then uses its new knowledge to choose the next query.

Choosing a query is similar to decision making under uncertainty: the most informative query is one that minimizes a loss function in expectation (Lindley 1956). Our loss function represents the uncertainty of our estimates. The most common measure of uncertainty, and the one used in this paper, is Shannon's entropy, which quantifies the amount of information missing if the value of a random variable is unknown (Shannon, 1948). The units of entropy are *bits*, which is the amount of entropy in a single 50-50 Bernoulli trial.

Because solving active-learning minimization problems is often computationally infeasible, researchers have developed heuristic querying strategies. For example *uncertainty sampling* selects that query about which the current estimate of the model is most uncertain (e.g., Lewis and Gale, 1994). Uncertainty sampling is equivalent to posterior entropy minimization under special conditions and approximates posterior entropy minimization for most problems. (Appendix 1 provides an example where it is equivalent.) Uncertainty sampling is known in marketing science as choice balance as used in both aggregate customization and choice-based polyhedral methods (e.g., Arora and Huber 2001; Toubia, Hauser and Simester 2004).

Decision Heuristics for Consideration Sets

When faced with many alternatives and/or many product features consumers make decisions by first forming a consideration set and then choosing from within the consideration set (e.g., Bronnenberg and Vanhonacker 1996; DeSarbo et al., 1996; Hauser and Wernerfelt 1990; Jedidi, Kohli and DeSarbo, 1996; Mehta, Rajiv and Srinivasan, 2003; Montgomery and Svenson 1976; Payne 1976; Roberts and Lattin, 1991; Shocker et al., 1991; Wu and Rangaswamy 2003). Consideration sets are often small compared to the set of available products – about 10% for common consumer goods and about 2-4% for complex products such as automobiles (Hauser and Wernerfelt 1990). Just knowing the consideration set can explain up to 80% of the uncertainty in choice models (Hauser 1978).

Prior research suggests that, to form consideration sets, consumers use heuristic decision rules that are likely less taxing than a full evaluation of expected utility (e.g., Bettman, Luce and Payne 1998; Bröder 2000; Gigerenzer and Goldstein 1996; Gigerenzer and Todd 1999; Hogarth and Karelaia 2005; Payne, Johnson and Bettman 1988, 1993; Martignon and Hoffrage 2002; Simon 1955; Shugan 1980). Researchers have proposed many decision heuristics including conjunctive, disjunctive, elimination-by-aspects, lexicographic, and subset conjunctive rules (e.g., Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005; Montgomery and Svenson 1976; Payne, Bettman and Johnson 1988; Tversky 1972; Yee, et al. 2007). Hauser, et al. (2009) demonstrate that, for consideration decisions, each of these rules can be written as a disjunction of conjunctive decision rules. Empirically, they find that roughly 93% of global-positioning-system consumers use a single conjunction and the remaining use two conjunctions. We begin by focusing on identifying conjunctive rules and return to the more-general model at the end of the paper. In a conjunctive rule, the consumer chooses a subset of feature-levels and, if the product satisfies those feature-levels, the consumer considers the product for further evaluation.

Inference of Decision Heuristics for Consideration-Set Formation

There are three categories of approaches for inferring consideration heuristics: (1) consideration and decision rules revealed as latent constructs, (2) consideration measured directly then decision rules revealed by the ability of the rules to fit observed consideration decisions, and (3) decision rules measured directly through self-explicated questions. See review in Ding, Hauser and Gaskin 2009. Categories (1) and (2) include maximum-likelihood, simulated-maximum likelihood, Bayesian, and machine-learning methods (Andrews and Srinivasan 1995;

Chiang, Chib and Narasimhan 1999; Erdem and Swait 2004; Gilbride and Allenby 2004, 2006; Hauser, et al. 2009; Swait and Ben-Akiva1987; Yee, et al. 2007). While successful in moderately-sized applications, these methods become impractical for complex product categories because computations grow exponentially with the number of feature levels. We seek to improve data collection and make estimation practical for approaches in the spirit of Categories (1) and (2).

Category (3) methods, directly stated decision rules, do not suffer the curse of dimensionality and might provide an alternative approach to decision-rule inference. While early attempts which asked consumers to self-state "unacceptable levels" led to many difficulties, more recent attempts are promising (Ding, et al. 2009; Green, Krieger and Banal 1988; Malhotra 1986; Klein 1986; Srinivasan 1988; Srinivasan and Wyner 1988; Sawtooth 1996). By improving the feasibility of decompositional methods for complex categories our research should enable future papers to compare decompositional methods to the compositional methods of Category (3).

One adaptive algorithm is now in common use. The first stage of Adaptive Choice-Based Conjoint Analysis (ACBC, Sawtooth 2008) asks consumers to evaluate approximately twenty-eight profiles that are variations on a "bring-your-own" profile. If the ACBC algorithm "notices" that some feature levels are <u>always</u> used conjunctively, it queries the consumer to confirm or disavow the suggested conjunctive rule. The current implementation uses rules of thumb, but appears to be a key component in making the choice-based stage of ACBC accurate. The algorithm developed in this paper can be used to make the ACBC queries close to optimal, to quantify the probability that a feature level is being used conjunctively, and to enable the first stage of ACBC to scale to much larger problems than are currently feasible.

3. Model of the Consumer Consideration Decision

In the conjunctive decision rule, a consumer considers a product or profile (hereafter profile) if all features of the profile have levels that are acceptable. (A feature becomes irrelevant to the decision rule if all of its levels are acceptable.) Because a profile either has a feature level or does not, we follow Tversky (1972) and refer to a feature level as an aspect. In this definition a binary feature, e.g., hybrid engine or not, is represented by two aspects. We discretize continuous features, such as miles per gallon (mpg), into mutually exclusive and collectively exhaustive aspects (e.g., 25-30 mpg). Using these definitions, a conjunctive decision rule implies that a profile is considered if and only if all of its aspects acceptable.

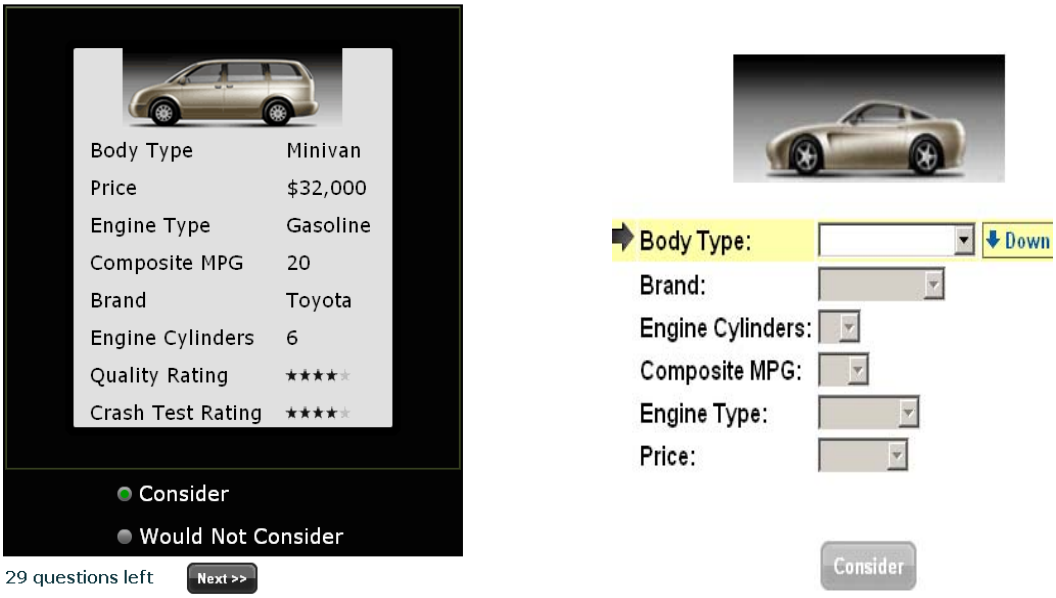
Let M be the number of features (e.g., brand, body style) and let N be the total numbers of

aspects (e.g., Audi, Honda, coupe, crash-test rating of 5, 25-30 mpg). Let *i* index consumers and *j* index aspects. Consumer *i*'s decision rule is a vector, $\vec{a}_i$, of length *N*, with elements a_{ij} such that $a_{ij} = 1$ if aspect *j* is acceptable and $a_{ij} = -1$ if it is not. Each query, indexed by *k*, is a profile, $\vec{x}_{ik}$, with *N* elements, x_{ijk}, such that $x_{ijk} = 1$ if profile *k* has aspect *j* and $x_{ijk} = 0$ if it does not. Each $\vec{x}_{ik}$ has exactly *M* non-zero elements, one for each feature (a profile contains one brand, one body type, one engine type, etc.).

A profile $\vec{x}_{ik}$ *satisfies* a conjunctive rule $\vec{a}_i$ if whenever $x_{ijk} = 1$, then $a_{ij} = 1$. This can be rewritten as $\min_j\{x_{ijk}a_{ij}\} \geq 0$. Following Gilbride and Allenby (2004), we define an indicator function such that $I(\vec{x}_{ik}, \vec{a}_i) = 1$ if $\min_j\{x_{ijk}a_{ij}\} \geq 0$, and $I(\vec{x}_{ik}, \vec{a}_i) = 0$ otherwise.

In a query a consumer is presented a profile and chooses a label, "consider" or "not consider." (We use the word "label" as in the active-learning literature.) See Figure 1(a) for an example query. If there were no response errors we would observe "consider" if and only if $I(\vec{x}_{ik}, \vec{a}_i) = 1$. But we allow response errors. Specifically, we assume that a consumer gives a false-positive answer with probability ϵ_1 and a false-negative answer with probability ϵ_2.

Figure 1
Example Queries and Example Configurator



(a) Example Query (b) Example Configurator

Let X_{iK} be the matrix of the first *K* profiles given to a consumer. Let y_{ik} be consumer *i*'s label for the k^{th} profile, where $y_{ik} = 1$ if the consumer says "consider" and $y_{ik} = 0$ otherwise.

Let $\vec{y}_{iK}$ be the vector of the first *K* labels. Then, for any potential conjunctive decision rule, $\vec{a}_i$, we have the following data-generating model:

$$\Pr(y_{ik} = 1 \mid \vec{x}_{ik}, \vec{a}_i) = (1 - \epsilon_1) I(\vec{x}_{ik}, \vec{a}_i) + \epsilon_1 \big(1 - I(\vec{x}_{ik}, \vec{a}_i)\big) \quad (1)$$

$$\Pr(y_{ik} = 0 \mid \vec{x}_{ik}, \vec{a}_i) = \epsilon_2 I(\vec{x}_{ik}, \vec{a}_i) + (1 - \epsilon_2)(1 - I(\vec{x}_{ik}, \vec{a}_i))$$

Once data are obtained, the inference problem is to compute a posterior distribution for the $\vec{a}_i$'s based on $\vec{y}_{iK}$ and X_k. Our goal is to select the $\vec{x}_{ik}$'s to get as much information as feasible about the $\vec{a}_i$'s.

4. Question Selection Using Active Learning

We propose the following framework for active-learning adaptive question selection:

1. Initialize beliefs by generating consumer-specific priors.
2. Use current beliefs to select the query that leads to greatest reduction in uncertainty in beliefs.
3. Query the consumer and update beliefs with the observed answer (label).
4. Continuing looping Steps 2 and 3 until a stopping rule is reached.

Initialize Consumer-Specific Beliefs

Empirically, a small fraction of possible profiles are considered by consumers. With non-informative priors, the initial queries are essentially random guesses, most of which will not be considered by the consumer. When a consumer considers a profile we learn (subject to the errors) that all of its aspects are acceptable; when a consumer rejects a profile we learn only that one or more aspects are unacceptable. Therefore, until the system observes the first considered profile, its queries are less efficient. That first "yes" provides substantial information.

Configurator. To initialize consumer-specific priors we ask the consumer to specify one vehicle that they will consider. Such tasks are often called "configurators" (Dahan and Hauser 2002). Based on pretests, we ask the consumer to configure only the horizontal features such as brand or body type. For vertical features, such as mpg, we assume the highest level is acceptable. A configurator is roughly the same task as ACBC's "bring your own" profile, which is justified on more intuitive grounds (Sawtooth 2008). While one might ask consumers to configure multiple vehicles, neither we nor Sawtooth find a second configurator particularly effective because later configured profiles tend to be similar to the first.

Some features, such as brand, have many aspects, but the configurator provides focused

information on only one brand. It is often these myriad-level features that represent the most heterogeneity in the market, e.g., consumers vary widely in the brands of automobiles they would consider. Because firms often have co-occurrence data on brand consideration (and other features), we use these data to focus initial beliefs about a consumer. Without loss of generality, let $j = 1$ index the configured brand aspect and, for other brand aspects, $j \neq 1$, let b_{1j} be the probability that $a_{ij} = 1$ when $a_{i1} = 1$. We set prior beliefs on the marginal probabilities for the brand aspects, j, to reflect the market-level co-occurrence data: $\Pr(a_{i1} = 1|\vec{x}_{i1}, y_{i1}) = 1$ and $\Pr(a_{ij} = 1|\vec{x}_{i1}, y_{i1}) = b_{ij}$. For configured non-brand aspects, j', we set $\Pr(a_{ij'} = 1|\vec{x}_{i1}, y_{i1}) = 1$ and use weakly informative priors for all other aspects. (It will become clear how we use these marginal probabilities when we discuss the belief-propagation algorithm.)

Pseudo-questions. We know from market-level data that knowing that some aspects are in a decision rule gives us information about whether other aspects are in that decision rule. For example, a consumer who will consider a Porsche is less likely to consider a Kia (brand) or a pick-up (body style). We include this market-level information in the priors for each consumer with a trick we call "pseudo-questions." For example, a pseudo-question appends a pseudo observation indicating that the consumer answered "not consider" to queries such as "Porsche ∩ Kia" or "Porsche ∩ pick-up." Within the model this implies that Kia and Porsche are not both considered. Because we do not know for certain "Porsche" will always rule out "Kia," or vice versa, we implement market-level priors by manipulating the ϵ's for pseudo-questions. For example, we set $\epsilon_2 = 0.15$ for "Porsche ∩ pick-up." This form of prior information is well-matched to belief-propagation algorithms and, empirically, seems to be efficient and effective.

Update Beliefs Based on Observed Responses

We use Bayes Theorem to update our beliefs after the K^{th} query:

$$\Pr(\vec{a}_i|X_{iK}, \vec{y}_{iK}) \propto \Pr(y_{iK}|\vec{x}_{iK}, \vec{a}_i = \vec{a}) \Pr(\vec{a} = \vec{a}_i|X_{i,K-1}, \vec{y}_{i,K-1}) \quad (2)$$

where we obtain $\Pr(y_{iK}|\vec{x}_{iK}, \vec{a}_i = \vec{a})$ from the data-generating model in Equation 1. The variable of interest, $\vec{a}_i$, is defined over all binary vectors of length N. With 53 aspects in our application, there are $2^{53} = 9.0 \times 10^{15}$ potential conjunctions of aspects. Without further structure we would need to update a discrete distribution defined on this space. This poses challenges to commonly used methods for sampling posterior distributions.

To gain insight for a feasible algorithm we examine solutions to related problems. Gil-

bride and Allenby (2004) use Ritter and Tanner's (1992) "Griddy Gibbs" algorithm to sample threshold levels for features. At the consumer level, the thresholds are drawn from a multinomial distribution. The Griddy Gibbs uses a grid approximation to the (often univariate) conditional posterior. We cannot modify their solution directly, in part because most of our features are horizontal (e.g., brand) and thresholds do not apply. Instead of estimating one value per feature (the threshold), we need to classify each level as acceptable or not.

We also examine the elimination-by-aspects (EBA) algorithm proposed by Gilbride and Allenby (2006). Their structure also does match our needs for complex products. Even if the structure could be modified, based on published applications, neither the threshold nor the EBA MCMC algorithms would run sufficiently fast between questions, even for small problems. If delays are more than a second or so, consumers find the delays annoying (about 0.4 seconds or better is best). Also, our problem is different in kind from either algorithm. We need to update beliefs between queries only for the current consumer. Population distributions need not be updated. But we can draw insight from the Gilbride-Allenby algorithms and use a binomial representation of aspect acceptance probabilities as applied at the level of the consumer.

For a feasible algorithm we estimate the binary aspect acceptance model with a variational Bayes approach. In variational Bayes inference, a complex posterior distribution is approximated with a variational distribution chosen from a family of posterior distributions similar to the true posterior distribution. Ideally, the variational family can be evaluated quickly (Attias 1999, Ghahramani and Beal 2000). Rapid computations are particularly important to choose consideration queries because a look forward requires two posterior updates for every potential query – once for a "yes" label and once for a "no" label. To reduce computational time from $o(2^N)$ to $o(N)$, we approximate the distribution of $\vec{a}_i$ as a vector of independent a_{ij}'s. For notational simplicity we define $p_{ijK} = \Pr\left(a_{ij} = 1 \middle| X_{iK}, \vec{y}_{iK}\right)$. Because this variational approximation is within a consumer, we place no restriction on the empirical population distribution of the a_{ij}'s. Intercorrelation at the population level is likely (and allowed) among aspect probabilities.

To calculate posteriors for the p_{ijK}'s we use a version of belief propagation (Yedidia, Freeman and Weiss 2003; Ghahramani and Beal 2001). The algorithm iteratively converges to an estimate of $\vec{p}_{iK}$. The h^{th} iteration uses Bayes Theorem to update each p_{ijK}^h based on the data and based on $p_{ij'K}^h$ for $j' \neq j$. Within the h^{th} iteration the algorithm loops over aspects and que-

ries using the data-generating model (Equation 1) to compute the likelihood of observing $y_k = 1$ conditioned on the likelihood for $k' \neq k$. In our experience, the algorithm converges quickly to posterior distributions that fit well the observed queries. Appendix 2 provides the pseudo-code.

Selecting the Next Question (Query)

Based on posterior beliefs, we select the query such that we learn as much as possible from the consumer's answer (label). Subject to computational issues, we select the K+1st query, $\vec{x}_{i,K+1}$, to minimize the posterior entropy that we expect after the consumer answers the query, $E[H(\vec{a}_i|X_{iK+1}, \vec{y}_{iK})]$.

To compute the expected posterior entropy, we first compute the probability of a "yes" label, $\Pr(y_{i,K+1} = 1|X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1})$, from our current beliefs about the probabilities for the decision rules. To evaluate a potential query, $\vec{x}_{i,K+1}$, we compute a posterior distribution for the decision rules conditioned on $y_{i,K+1} = 1$ and we compute a posterior distribution conditioned on $y_{i,K+1} = 0$. The expected posterior entropy is the expectation over the probabilities of "yes" and "no" answers to the query. We compute the entropy of our estimate of the a_{ij}'s. Specifically:

$$(3)\quad E\left[H\left(\vec{x}_{i,K+1}|X_{iK}, \vec{y}_{iK}\right)\right] = -q_{i,K+1}\left(\vec{x}_{i,K+1}\right)\sum_j \left\{ \begin{matrix} p^{+}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\log_2\left[p^{+}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\right] + \\ \left[1 - p^{+}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\right]\log_2\left[1 - p^{+}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\right] \end{matrix} \right\}$$
$$-\left[1 - q_{i,K+1}\left(\vec{x}_{i,K+1}\right)\right]\sum_j \left\{ \begin{matrix} p^{-}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\log_2\left[p^{-}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\right] + \\ \left[1 - p^{-}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\right]\log_2\left[1 - p^{-}_{ij,K+1}\left(\vec{x}_{i,K+1}\right)\right] \end{matrix} \right\}$$

where, $p^{+}_{ij,K+1}(\vec{x}_{i,K+1}) = \Pr(a_{ij} = 1|X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1}, y_{i,K+1} = 1)$, $p^{-}_{ij,K+1}(\vec{x}_{i,K+1}) = \Pr(a_{ij} = 1|X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1}, y_{i,K+1} = -1)$, and $q_{i,K+1}\left(\vec{x}_{i,K+1}\right) = \Pr(y_{i,K+1} = 1|X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1})$.

For small and moderate empirical problems we evaluate the expected posterior entropy for every potential query, but for complex products such as automobiles this is not feasible; there are 357,210 potential queries in our application. Even after eliminating unrealistic profiles, such as a Lexus pick-up truck, there are 136,144 potential queries. Instead, we combine posterior entropy minimization with uncertainty sampling.

Our modification is simple; we select the T queries about which we are most uncertain ($q_{i,K+1}\left(\vec{x}_{i,K+1}\right) \approx 0.5$), then choose the query which minimizes posterior entropy among the T queries. (Both theoretically and empirically, the most choice-balanced queries do not necessarily minimize expected posterior entropy.) We select T to balance performance and speed. For example, in simulations with T equal to 64 queries out of all 256 potential profiles, we get 95% of

the optimal posterior entropy 89.9% of the time. The automobile application used $T = 1000$.

Stopping Rules

In our application, based on extensive experience in the product category, automotive managers with market-research experience judged that we should stop after 30 calibration queries. Fortunately, the 30-profile managerial decision does not affect the basic insights we obtain from our research. Future research might collect new data to explore other stopping rules, for example, rules that balance the cognitive cost to the consumer of more queries with the potential reduction in uncertainty to the manager.

5. Synthetic-Data Experiments

To evaluate the active-learning algorithm, we choose a synthetic problem for which both (D-efficient) orthogonal designs and random designs are expected to do reasonably well. With four features at four levels each (4^4) an orthogonal design is 32 profiles and we can randomly cycle through all 256 profiles to illustrate comparable learning curves.

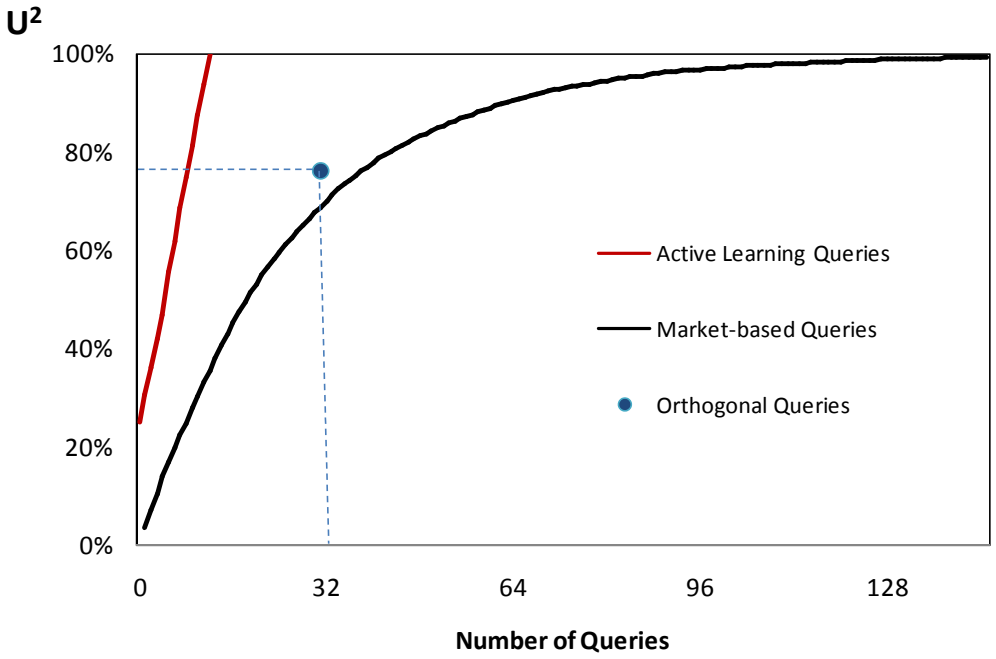
The 16 aspects imply 65,536 potential decision rules from which we randomly select 1,000 synthetic respondents. Each synthetic respondent answers queries generated by orthogonal, random, market-based, and active-learning designs.[1] We compute posterior decision-rule probabilities based on $K = 1$ to 256 queries for random, market-based, and active-learning questions. An orthogonal design is fixed at 32 queries by definition. Market-based queries are chosen randomly but in proportion to the market share that profiles might expect in the market (the market shares are known for synthetic respondents).

In synthetic data we know the decision rule as a binary vector and we predict the probabilities that each value is 0 vs. 1. The appropriate statistic is the percent of uncertainty explained (U^2, Hauser 1978).[2] Key results are presented in Figure 2. To simplify interpretation we plot random queries in Appendix 4 rather than Figure 2, because the results are indistinguishable from random queries on the scale of Figure 2. Market-based queries do about 3% better than random queries for the first 16 queries, about 1% better for the first 32 queries, and about ½% better for all 256 queries. (Queries 129 through 256 are not shown in Figure 2; the random-query and the market-based query curves asymptote to 100%.)

[1] For clarity with synthetic data, we set $\epsilon_1 = \epsilon_2 = 0$ and $T = 256$. Qualitative implications and relative comparisons are similar with error-laden responses ($\epsilon_1, \epsilon_2 > 0$) and with uncertainty sampling ($T < 256$).

[2] In the synthetic data the decision rule for each aspect is probabilistically independent, hence it is appropriate to define U^2 on aspect-level decision rules.

Figure 2
Percent Uncertainty Explained (U^2) for Alternative Querying Methods



Active-learning queries do well (red line). For a given level of accuracy, active-learning queries require comparatively fewer questions. For example, an orthogonal design achieves a U^2 of approximately 76% in 32 queries. To get the same U^2 requires 38 market-based queries or 40 random queries, but only 9 active-learning queries. The improvement comes both from efficient active-learning queries after the configurator and from the configurator. To examine the influence of the configurator alone, Appendix 4 plots a mixed model of a configurator plus random questions. The plot parallels the plot of purely random queries achieving a U^2 of approximately 76% in 30 queries.

Figure 2 also suggests that the belief-propagation algorithm achieves very close to an optimal reduction in posterior entropy on errorless data (review footnote 1). With 16 aspects and equally-likely priors, the prior entropy is 16 $\log_2(2)$, which is 16 bits. The configurator reveals four acceptable aspects: 4 bits. Each subsequent query is a Bernoulli trial that can reveal at most 1 bit, thus, a perfect active-learning algorithm would require 12 additional queries to identify a decision rule (4 bits + 12 bits identifies the 16 elements of $\vec{a}_i$). Figure 2 suggests we get very close to that theoretical maximum.

6. Illustrative Empirical Application: Vehicle Consideration

In Spring 2009 General Motors (GM) recognized that consideration of their vehicles was

well below that of non-US vehicles. Management was interested in exploring various means to increase consideration. As part of that effort, GM fielded a web-based survey to 2,336 respondents recruited and demographically-balanced from an automotive panel maintained by Harris Interactive, Inc. Respondents were screened to be 18 years of age and interested in purchasing a new vehicle in the next two years. Respondents received 300 Harris points (good for prizes) as compensation for completing a 40 minute survey. The response rate was 68.2% and the completion rate was 94.9%.

The bulk of GM's survey explored various marketing strategies that GM might use to enhance consideration of their brands, strategies such as communicating excellent J.D. Power or *Consumer Reports* ratings, projecting an image of cutting-edge development, and/or telling consumers that a GM brand is the top brand in key countries (Buick is #1 in China). Strategies were based on careful exploration and refinement through six months of qualitative interviews and concept tests. The managerial application is tangential to the scope and focus of this paper, but we can indicate that GM was able to identify strategies to increase consideration significantly relative to traditional automotive advertising. GM was also able to identify clusters of consumers who used similar heuristic decision rules and they are developing new strategies to address key clusters.

Because GM's future actions depend upon the accuracy with which they could evaluate whether their strategies and tactics affected decision rules, we were given the opportunity to test active-learning query-selection for a subset of the respondents. This subset of 872 respondents was not shown any induction. Instead, after configuring a profile, evaluating 29 calibration profiles, and completing a memory-cleansing task (Frederick 2005), respondents evaluated a second set of 29 validation profiles. (A 30th profile in calibration and validation was used for other research purposes by GM.) The profiles varied on 53 aspects: brand (21 aspects), body style (9 aspects), price (7 aspects), engine power (3 aspects), engine type (2 aspects), fuel efficiency (5 aspects), quality (3 aspects), and crash-test safety (3 aspects).

Calibration Profiles (Queries)

To test active-learning, one-half of the calibration queries were chosen by the active-learning algorithm. The other half were chosen randomly in proportion to market share from the top 50 best-selling vehicles in the US. To avoid order effects and to introduce variation in the data, the query-selection strategy was randomized. This probabilistic variation means that the

number of queries of each type is 14.5 on average, but varies by respondent.

We chose for comparison market-based rather than purely random queries based on the synthetic-data experiments. The market-based queries perform slightly better than purely-random queries and, hence, provide a stronger test. We could not test an orthogonal design because 29 queries is but a small fraction of the 13,320 profiles in an orthogonal design. Furthermore, even if we were to complete an orthogonal design of 13,320 queries, Figure 2 suggests that orthogonal queries do only slightly better than random or market-based queries.

Besides enabling methodological comparisons, this mix of adaptive and market-based queries has practical advantages with human respondents. First, the market-based queries introduce variety to engage the respondent and help disguise the choice-balance nature of the active-learning algorithm. (Respondents get variety in the profiles they evaluate.) Second, market-based queries sample "far away" from the active-learning queries and prevent the algorithm from getting stuck in a local maximum. Such far-away sampling is conceptually similar to exploration methods in simulated annealing and Metropolis-Hastings MCMC sampling.

Validation Profiles

Respondents were shown 29 market-based profiles. Because there was some overlap between the market-based validation and the market-based calibration profiles, we have an indicator of respondent reliability. Respondents consistently evaluated market-based profiles 90.5% of the time. Respondents are consistent, but not perfect, and, thus, modeling response error (via the ϵ's) appears to be appropriate. Furthermore, 90.5% might be an upper bound in terms of a predictive hit rate among non-configured profiles.

Performance Measures

Hit rate is an intuitive measure with which to compare predictive accuracy, but hit rate must be interpreted with caution for consideration data. If a respondent were to consider 20% of both calibration and validation profiles, then a null model that predicts "reject all profiles" will achieve a hit rate of 80%. But such a model provides little information, has a large number of false negative predictions, and predicts a consideration-set size of zero. On the other hand, a null model, that predicts randomly proportional to the consideration-set size in the calibration data, would predict a larger validation consideration-set size and balance false positives and false negatives, but would achieve a lower hit rate (68%: $0.68 = (0.8)^2 + (0.2)^2$). With this in mind, we report additional performance measures to help diagnose the relative performance of the active-

learning algorithm.

Besides hit rate, we examine false positive and false negative predictions. A manager might put more (or less) weight on not missing considered profiles than on predicting as considered profiles that are not considered. Without knowing specific loss functions to weigh false positives and false negatives differently, we use the Kullback-Leibler divergence (KL). KL divergence is a non-symmetric measure of the difference (in bits) from a prediction model to a comparison model (Chaloner and Verdinelli 1995; Kullback and Leibler 1951). It discriminates among models even when the hit rates might otherwise be equal. Appendix 3 provides formulae for the KL divergence measure appropriate to the data in this paper.

Empirical Results

Table 1 summarizes the four performance measures: hit rate (large is better), KL divergence (small is better), false positive percentage (small is better), and false negative percentage (small is better). All models are estimated with the variational-Bayes belief-propagation algorithm based only on the calibration data. We contrast the conjunctive model trained with the active-learning queries only, the market-based queries only, and based on both sets of queries combined. Table 1 also reports predictions for null models that predict all profiles as considered, predict no profiles as considered, and predict profiles randomly based on the consideration-set size among the calibration profiles. Due in part to the large sample size, all differences between the four models and all differences vs. the null models are significant ($p < 0.001$) with the exception of hit rates between the active-learning and the all-queries models ($p = 0.276$).

The active-learning algorithm improves predictions relative to market-based queries, doing significantly better on all measures. Naturally more data are better; performance measures improve when we add the market-based queries to the active-learning queries. Had GM collected data with 29 active-learning queries we hypothesize that predictions would improve further. A model based on approximately 14.5 active-learning queries per respondent does as well on hit rate as a model based on all 29 queries and almost as well on the other measures.

The nine-month managerial study asked respondents to evaluate profiles before and after an induction – a total of 60 profiles. (Managerial goals constrained the architecture of the validation study in this paper.) Pretests indicated 60 profiles was close to the practical limit. Perhaps in the future data might become available in which all queries (or a larger percentage of queries) could be chosen adaptively.

Table 1
Performance of Query-Selection Methods (and other Models)
(small is better for KL divergence and for false positives and false negatives)*

	Hit Rate	KL Divergence	Percent False Positive	Percent False Negative
Active-learning queries (~14.5 queries)	84.6%	0.475	3.1%	12.3%
Market-based queries (~14.5 queries)	82.7%	0.512	2.3%	14.9%
All queries (29 queries)	84.8%	0.452	3.7%	11.5%
HB Additive Partworths (29 queries)	81.5%	0.561	3.3%	15.3%
Consider all (null)	18.0%	0.565	82.0%	0.0%
Consider none (null)	82.0%	0.565	0.0%	18.0%
Random predictions (null)	73.2%	0.562	13.8%	13.0%

*All comparisons are significant at the $p < 0.001$ level, except the comparison between active-learning queries and all queries on hit rate.If we were to estimate an additive-partworth model, the common standard would be hierarchical Bayes (HB) estimation. However, with 53 aspects such a model would typically rely on more observations than the 29 in our data. Nonetheless, it is interesting to report the predictions we obtain from such a model. An HB additive model on all the data does not do as well as any of the heuristic-decision-rule models. We interpret these results as evidence that a non-compensatory decision rules fit the data reasonably well, not that we can reject compensatory models. HB-additive is known to predict well when the ratio of observations to parameters for each respondent is not as small. HB-additive might do much better if we had data on substantially more queries per respondent.

Summary of Empirical Illustration

Active-learning queries are promising. We appear to be able to select queries to provide significantly more information per query than market-based queries, which, in turn, are comparable in synthetic data to random or orthogonal queries. Furthermore, with a relatively few queries (~14.5) we can do well even in a complex product category with 53 aspects.

6. Disjunctions of Conjunctive Decision Rules

The GM data focused on the consumer's primary conjunctive decision rule under the hypothesis that, within the limited number of queries that could be made, it was most managerially

important to identify the primary conjunction. In GPSs, a single conjunction was used by most consumers (~93%, Hauser, et al. 2009). But Table 1 hints that more data might identify more conjunctive rules (more queries reduce false negatives). Fortunately, our algorithm can be modified to search for a second conjunction and thus identify more-general heuristic decision rules. To do so, we would impose a dynamic stopping rule. Once a minimum calibration KL divergence is achieved, we would ask a second configurator pretested carefully to seek a maximally-different profile. We would re-initialize priors and restart the algorithm. If GM had focused all 60 queries on calibration, rather than just 29 queries, such a strategy would have been feasible.

As an initial indicator of the promise of such a strategy, we tried (1) estimating the best model with the data, (2) eliminating all calibration profiles that were correctly classified with the first conjunction, and (3) using the remaining market-based profiles to search for a second conjunction. As expected, this strategy reduced false negatives (10.6%), increased false positives slightly (4.9%), and maintained the hit rate (84.5%). We hope it inspires future applications that focus data collection on the empirical challenge of finding second conjunctions.

7. Other Applications of an Active-Learning Framework

Our focus was driven by an important managerial issue and an interesting theoretical challenge. However, the active-learning framework is not limited to estimating heuristic decision rules for consideration-set formation. All four steps apply to problems in which we seek to identify a binary performance vector (X_{iK}'s → $\vec{p}_i$'s). We suggest a few. In each of these applications the binomial approximation applies within a consumer, but there is no restriction on correlation at the population level among elements of the vector.

Product Design. Product-design engineers often employ decision matrices in which product features or engineering subsystems are varied systematically (e.g., Pugh 1991). Alternatively, product "clinics" present consumers with product concepts on which core benefits are varied systematically. In either case the elements can be selected with the proposed active-learning framework.

Store Layout. Store layout is a complex decision in which products are assigned to various regions of the store. Many retailers use (D-efficient) orthogonal experiments because changes are costly to implement and impose substantial opportunity costs. Store layout can be written as a binary assignment vector of products to regions and experiments might be made more efficient with active learning.

Integrated Marketing Communications. The number of media available to advertisers has exploded to include search engines, banner advertising, viral videos, social networks, and community enhancement as well as traditional media such as television, radio, magazines, newspapers, direct mail, and outdoor advertising. There is a long tradition of adaptive control of spending by media (Little 1966, 1977), but there is also the challenge of allocation across many discrete media. When the decision problem is binary (allocate to a media or not), the decision problem matches the proposed active-learning framework.

8. Open Questions and Challenges

In this paper we develop and test an active-learning algorithm to select queries when estimating non-compensatory decision rules for consideration decisions. To address applications in complex product categories we develop a variational-Bayes belief-propagation algorithm to update posteriors in real time. We select queries to minimize expect posterior entropy. Both synthetic-data experiments and the empirical illustration suggest that such queries provide significantly more information per query than extant methods.

We feel that our set of comparison methods represents question-selection practice and our use of belief-propagation variation-Bayes estimation is appropriate for the GM data. But there are opportunities for future research. Re-estimating decision rules from adaptive data is consistent with the likelihood principle, but currently-available non-compensatory methods are impractical for our application because of computational requirements. Fortunately, there is no reason to believe re-estimation would change the basic insight that (near) optimally-chosen queries provide improved information. Our results for discrete decision heuristics are consistent with research for continuous additive-partworth models where adaptive questions provide more information per question than orthogonal designs even when the models are re-estimated with hierarchical Bayes methods (e.g., Toubia, et al. 2003, but there are many other examples).

To make the active-learning algorithm feasible we made assumptions and selected "tuning" parameters by managerial judgment (see Little 1970). Standard leave-one-out cross-validation was not possible because the tuning parameters were selected prior to data collection. Because managerial judgment is never perfect, our results might improve with other choices of the ϵ's or T. Predictive ability might also improve with more data. Other interesting challenges include methods to distinguish between non-compensatory and compensatory decision rules and empirical comparisons to self-explicated decision rules.

References

Andrews, Rick L. and T. C. Srinivasan (1995), "Studying Consideration Effects in Empirical Choice Models Using Scanner Panel Data," *Journal of Marketing Research*, 32 (February), 30-41.

Ansari, Asim, Skander Essegaier and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37, 3, (August), 363-375.

------ and Carl F. Mela (2003), "E-Customization," *Journal of Marketing Research*, 40, (May), 131-145.

Attias, Hagai (1999), "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes," *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*.

Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28, (September), 273-283.

Bettman, James R., Mary Frances Luce and John W. Payne (1998), "Constructive Consumer Choice Processes," *Journal of Consumer Research*, 25, 3 (December), 187-217.

Bodapati, Anand V. (2008), "Recommendation Systems with Purchase Data," *Journal of Marketing Research*, 45, 1, (February), 77-93.

Bröder, Arndt (2000), "Assessing the Empirical Validity of the "Take the Best" Heuristic as a Model of Human Probabilistic Inference," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 26, 5, 1332-1346.

Bronnenberg, Bart J. and Wilfried R. Vanhonacker (1996), "Limited Choice Sets, Local Price Response, and Implied Measures of Price Competition," *Journal of Marketing Research*, 33 (May), 163-173.

Chaloner, Kathryn and Isabella Verdinelli (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 3, 273-304. (1995)

Chiang, Jeongwen, Siddhartha Chib and Chakravarthi Narasimhan (1999), "Markov Chain Monte Carlo and Models of Consideration Set and Parameter Heterogeneity," *Journal of Econometrics*, 89, 223-48.

Dahan, Ely and John R. Hauser (2002), "The Virtual Customer," *Journal of Product Innovation Management*, 19, 5, (September), 332-354.

Ding, Min, Steven Gaskin and John Hauser (2009), "A Critical Review of Non-compensatory and Compensatory Models of Consideration-Set Decisions," *2009 Sawtooth Software Conference Proceedings*, Delray, FL, March 23-27, 2009, 207-232.

------, John R. Hauser, Songting Dong, Daria Dzyabura, Zhilin Yang, Chenting Su and Steven Gaskin (2009), "Unstructured Direct Elicitation of Non-Compensatory and Compensatory Decision Rules," Working Paper, MIT Sloan School of Management.

DeSarbo, Wayne S., Donald R. Lehmann, Gregory Carpenter and Indrajit Sinha (1996), "A Stochastic Multidimensional Unfolding Approach for Representing Phased Decision Outcomes," *Psychometrika*, 61(3), 485-508.

Erdem, Tülin and Joffre Swait (2004), "Brand Credibility, Brand Consideration, and Choice," *Journal of Consumer Research*, 31 (June), 191-98.

Evgeniou, Theodoros, Constantinos Boussios and Giorgos Zacharia (2005), "Generalized Robust Conjoint Estimation," *Marketing Science*, 24(3), 415-429.

------, Massimiliano Pontil and Olivier Toubia (2007), "A Convex Optimization Approach to Modeling Heterogeneity in Conjoint Estimation," *Marketing Science*, 26, 6, (November-December), 805-818.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philos. Trans. Royal Society of London, Series A*, 222, 309-368.

Frederick, Shane (2005), "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives*. 19(4). 25-42.

Ghahramani, Zoubin and Beal, Matthew J (2000), "Variational Inference for Bayesian Mixtures of Factor Analysers." *Advances in Neural Information Processing Systems,* 12 (Cambridge, MA: MIT Press).

------ and ------ (2001), "Propagation Algorithms for Variational Bayesian Learning," *Advances in Neural Information Processing Systems*, 13 (Cambridge, MA: MIT Press).

Green, Paul E., Abba M. Krieger and Pradeep Bansal (1988), "Completely Unacceptable Levels in Conjoint Analysis: A Cautionary Note," *Journal of Marketing Research*, 25 (August), 293-300.

Gigerenzer, Gerd and Daniel G. Goldstein (1996), "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," *Psychological Review*, 103(4), 650-669.

------, Peter M. Todd and the ABC Research Group (1999), *Simple Heuristics That Make Us*

Smart, (Oxford, UK: Oxford University Press).

Gilbride, Timothy J. and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23(3), 391-406.

------ and ------ (2006), "Estimating Heterogeneous EBA and Economic Screening Rule Choice Models," *Marketing Science*, 25, 5, (September-October), 494-509.

Hauser, John R. (1978), "Testing the Accuracy, Usefulness and Significance of Probabilistic Models: An Information Theoretic Approach," *Operations Research*, Vol. 26, No. 3, (May-June), 406-421.

------, Olivier Toubia, Theodoros Evgeniou, Rene Befurt and Daria Dzyabura (2009), "Cognitive Simplicity and Consideration Sets," forthcoming, *Journal of Marketing Research*.

------ and Birger Wernerfelt (1990), "An Evaluation Cost Model of Consideration Sets," *Journal of Consumer Research*, 16 (March), 393-408.

Hogarth, Robin M. and Natalia Karelaia (2005), "Simple Models for Multiattribute Choice with Many Alternatives: When It Does and Does Not Pay to Face Trade-offs with Binary Attributes," *Management Science*, 51, 12, (December), 1860-1872.

Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33 (August), 307-317.

Jedidi, Kamel and Rajeev Kohli (2005), "Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers," *Journal of Marketing Research*, 42 (4), 483-494.

------, ------ and Wayne S. DeSarbo (1996), "Consideration Sets in Conjoint Analysis," *Journal of Marketing Research*, 33 (August), 364-372.

Johnson, Richard (1987), "Accuracy of Utility Estimation in ACA," Working Paper, Sawtooth Software, Sequim, WA, (April).

---- (1991), "Comment on `Adaptive Conjoint Analysis: Some Caveats and Suggestions,'" *Journal of Marketing Research*, 28, (May), 223-225.

Klein, Noreen M. (1988), "Assessing Unacceptable Attribute Levels in Conjoint Analysis," *Advances in Consumer Research* vol. XIV, pp. 154-158.

Lewis, David D. and William A Gale (1994), "Training Text Classifiers by Uncertainty Sampling," *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 3-12.

Lindley, D. V. (1956), "On a Measure of the Information Provided by an Experiment," *The An-*

nals of Mathematical Statistics, 27, 986-1005.

Little, John D. C. (1966), "A Model of Adaptive Control of Promotional Spending," *Operations Research*, 14(6) 1075-1097.

------ (1970), "Managers and Models: The Concept of a Decision Calculus," *Management Science*, 16 (8) B466-485.

------ (1977), "Optimal Adaptive Control: A Multivariate Model for Marketing Applications," *IEEE Transactions on Automatic Control*, 22(2) 187-195.

Liu, Qing, Thomas Otter and Greg M. Allenby (2007), "Investigating Endogeneity Bias in Marketing," *Marketing Science*, 26, 5, (September-October), 642-650.

Malhotra, Naresh (1986), "An Approach to the Measurement of Consumer Preferences Using Limited Information," *Journal of Marketing Research*, 23 (February), 33-40.

Martignon, Laura and Ulrich Hoffrage (2002), "Fast, Frugal, and Fit: Simple Heuristics for Paired Comparisons," *Theory and Decision*, 52, 29-71.

Mehta, Nitin, Surendra Rajiv and Kannan Srinivasan (2003), "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation," *Marketing Science*, 22(1), 58-84.

Montgomery, Alan L., Shibo Li, Kannan Srinivasan and John Liechty (2004), "Modeling Online Browsing and Path Analysis Using Clickstream Data," *Marketing Science*, 23, 4, (Fall), 579-585.

Montgomery, H. and O. Svenson (1976), "On Decision Rules and Information Processing Strategies for Choices among Multiattribute Alternatives," *Scandinavian Journal of Psychology*, 17, 283-291.

Netzer, Oded and V. Srinivasan (2007), "Adaptive Self-Explication of Multi-Attribute Preferences," (Palo Alto, CA; Stanford University), Working Paper.

Ordóñez, Lisa D., Lehmann Benson III and Lee Roy Beach (1999), "Testing the Compatibility Test: How Instructions, Accountability, and Anticipated Regret Affect Prechoice Screening of Options," *Organizational Behavior and Human Decision Processes*, 78, 1, (April), 63-80.

Payne, John W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search," *Organizational Behavior and Human Performance*, 16, 366-387.

------, James R. Bettman and Eric J. Johnson (1988), "Adaptive Strategy Selection in Decision

Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 534-552.

------, ------ and ------ (1993), *The Adaptive Decision Maker*, (Cambridge UK: Cambridge University Press)..

Pugh, Stuart (1991). *Total Design: Integrated Methods for Successful Product Engineering*, (Reading, MA: Addison-Wesley).

Ritter, Christian and Martin A. Tanner (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 419, 861-868.

Roberts, John H. and James M. Lattin (1991), "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28 (November), 429-440.

Sandor, Zsolt and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," *Journal of Marketing Research*, 38, 4, (November), 430-444.

---- and ---- (2002), "Profile Construction in Experimental Choice Designs for Mixed Logit Models," *Marketing Science*, 21, 4, (Fall), 398–411.

Sawtooth Software, Inc. (1996), "ACA System: Adaptive Conjoint Analysis," ACA Manual, (Sequim, WA: Sawtooth Software, Inc.)

------ (2008), "ACBC Technical Paper," (Sequim WA; Sawtooth Software, Inc.)

Settles (2009), "Active Learning Literature Survey," (Madison, WI: University of Wisconsin – Madison), Computer Sciences Technical Report 1648.

Shannon, C.E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, (July & October) 379–423 & 623–656.

Shocker, Allan D., Moshe Ben-Akiva, Bruno Boccara, and Prakash Nedungadi (1991), "Consideration Set Influences on Consumer Decision-Making and Choice: Issues, Models, and Suggestions," *Marketing Letters*, 2(3), 181-197.

Shugan, Steven (1980), "The Cost of Thinking," *Journal of Consumer Research*, 27(2), 99-111.

Simester, Duncan I., Peng Sun and John N. Tsitsiklis (2006), "Dynamic Catalog Mailing Policies," *Management Science*, 52, 5, (May), 683-696.

Simon, Herbert A. (1955), "A Behavioral Model of Rational Choice," *The Quarterly Journal of Economics*, 69(1). 99-118.

Srinivasan, V. (1988), "A Conjunctive-Compensatory Approach to The Self-Explication of Mul-

tiattributed Preferences," *Decision Sciences*, 295-305.

------ and Gordon A. Wyner (1988), "Casemap: Computer-Assisted Self-Explication of Multiattributed Preferences," in W. Henry, M. Menasco and K. Takada, Eds, *Handbook on New Product Development and Testing*, (Lexington, MA: D. C. Heath), 91-112.

Swait, Joffre and Tülin Erdem (2007), "Brand Effects on Choice and Choice Set Formation Under Uncertainty," *Marketing Science* 26, 5, (September-October), 679-697.

Toubia, Olivier, John R. Hauser and Rosanna Garcia (2007), "Probabilistic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis: Theory and Application," *Marketing Science*, 26, 5, (September-October), 596-610..

------, ------ and Duncan I. Simester (2004), "Polyhedral Methods for Adaptive Choice-based Conjoint Analysis," *Journal of Marketing Research*, 41, (February), 116-131.

------, Duncan I. Simester, John R. Hauser and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22, 3, (Summer), 273-303.

Tversky, Amos (1972), "Elimination by Aspects: a Theory of Choice," *Psychological Review*, 79(4), 281-299.

Glen L. Urban and John R. Hauser (2004), "'Listening-In' to Find and Explore New Combinations of Customer Needs," *Journal of Marketing*, 68, (April), 72-87.

Wu, Jianan and Arvind Rangaswamy (2003), "A Fuzzy Set Model of Search and Consideration with an Application to an Online Market," *Marketing Science*, 22(3), 411-434.

Yedidia, Jonathan S., William T. Freeman and Yair Weiss (2003), *Understanding Belief Propagation and Its Generalizations: Exploring Artificial Intelligence in The New Millennium*, (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 239-269.

Yee, Michael, Ely Dahan, John R. Hauser and James Orlin (2007) "Greedoid-Based Noncompensatory Inference," *Marketing Science*, 26, 4, (July-August), 532-549.

Ying, Yuanping, Fred Feinberg and Michel Wedel (2006), "Leveraging Missing Ratings to Improve Online Recommendation Systems," *Journal of Marketing Research*, 43, 3, (August), 355-365.

Appendix 1. Example Where Uncertainty Sampling Minimizes Posterior Entropy

We choose a simple example with two aspects to demonstrate the intuition. This example generalizes to *N* aspects. We abstract away from response error by setting $\epsilon_1 = \epsilon_2 = 0$ and we choose uninformative priors such that $p_{i1}^0 = p_{i2}^0 = 0.5$. With two aspects there are four potential queries, $\vec{x}_{i1}$ = {0, 0}, {0, 1}, {1, 0}, and {1, 1}, and four potential decision rules, $\vec{a}_i$ = {0, 0}, {0, 1}, {1, 0}, and {1, 1}, each of which is *a priori* equally likely. But the different $\vec{x}_{i1}$'s provide differential information about the decision rules. For example, if $\vec{x}_{i1}$ = {0, 0} and $y_{i1} = 1$ then the decision rule must be $\vec{a}_i$ = {0, 0}. At the other extreme, if $\vec{x}_{i1}$ = {1, 1} and $y_{i1} = 1$, then all decision rules are consistent. The other two profiles are each consistent with half of the decision rules. We compute $\Pr(y_{i1} = 1 \mid \vec{x}_{i1})$ for the four potential queries as 0.25, 0.50, 0.50, and 1.00, respectively.

We use the formulae in the text for expected posterior entropy, $E[H(\vec{x}_{i1})]$:

Potential Query ($\vec{x}_{i1}$)	Pr ($y_{i1} = 1 \mid \vec{x}_{i1}$)	$E[H(\vec{x}_{i1})]$
{0, 0}	0.25	$-\frac{3}{2}\left(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}\right) = 1.4$
{0, 1}	0.50	$-\log_2\frac{1}{2} = 1$
{1, 0}	0.50	$-\log_2\frac{1}{2} = 1$
{1, 1}	1.00	$-2\log_2\frac{1}{2} = 2$

By example expected posterior entropy is minimized for either of the queries, {0, 1} or {1, 0}, both of which are consistent with uncertainty sampling (choice balance).

Appendix 2. Pseudo-Code for Belief Propagation Algorithm

Maintain the notation of the text and let $\vec{p}_{iK}$ be the vector of the p_{ijK}'s, let $\vec{p}_{iK,-j}$ be the vector of all but the j^{th} element, let $\vec{y}_{K,-k}$ be the vector of all labels except the k^{th} label, and define two index sets, $S_j^+ = \{k | x_{ijk} = 1, y_k = 1\}$ and $S_j^- = \{k | x_{ijk} = 1, y_k = 0\}$. Let superscript *h* index an iteration with *h* = 0 indicating a prior. The belief-propagation algorithm uses all of the data, X_K and $\vec{y}_{iK}$, when updating for the K^{th} query. In application, the ϵ's are set by managerial judgment prior to data collection. Our application used $\epsilon_1 = \epsilon_2 = 0.01$ for query selection.

Use the priors to initialize $\vec{p}_{iK}^{0}$. Initialize all $\Pr(y_{iK,-k}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=\pm 1)$.

While $\max_j(p_{ijK}^{h}-p_{ijK}^{h-1})>0.001$.

For j = 1 to N

For $k\in S_j^+$

$$\Pr(y_k=1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)=(1-\epsilon_1)\prod_{x_{igk}=1,g\neq j}p_{igK}^{h-1}+\epsilon_1(1-\prod_{x_{igk}=1,g\neq j}p_{igK}^{h-1})$$

$$\Pr(y_k=1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)=(1-\epsilon_1)$$

$$\Pr(\vec{y}_K|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)=\Pr(\vec{y}_{K,-k}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)\Pr(y_k=1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)$$

$$\Pr(\vec{y}_K|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)=\Pr(\vec{y}_{K,-k}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)\Pr(y_k=1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)$$

where $\Pr(\vec{y}_{K,-k}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=\pm 1)$ is computed as the product of marginals

end loop $k\in S_j^+$

For $k\in S_j^-$

$$\Pr(y_k=-1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)=(1-\epsilon_1)(1-\prod_{x_{igk}=1,g\neq j}p_{igK}^{h-1})+\epsilon_1\prod_{x_{igk}=1,g\neq j}p_{igK}^{h-1}$$

$$\Pr(y_k=-1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)=(1-\epsilon_1)$$

$$\Pr(\vec{y}_K|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)=\Pr(\vec{y}_{K,-k}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)\Pr(y_k=-1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)$$

$$\Pr(\vec{y}_K|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)=\Pr(\vec{y}_{K,-k}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)\Pr(y_k=-1|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)$$

end loop $k\in S_j^-$

$$\Pr(a_{ij}=1|X_{iK},\vec{y}_{iK},\vec{p}_{iK,-j}^{h-1})\propto\Pr(\vec{y}_{iK}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=1)\Pr(a_{ij}=1|prior)$$

$$\Pr(a_{ij}=-1|X_{iK},\vec{y}_{iK},\vec{p}_{iK,-j}^{h-1})\propto\Pr(\vec{y}_{iK}|X_{iK},\vec{p}_{iK,-j}^{h-1},a_{ij}=-1)(1-\Pr(a_{ij}=1|prior))$$

$p_{ijK}^{h}=\Pr(a_{ij}=1|X_{iK},\vec{y}_{iK},\vec{p}_{iK,-j}^{h-1})$ normalized.

end loop j

Appendix 3. Kullback-Leibler Divergence for Our Data

The Kullback-Leibler divergence (KL) is an information-theory-based measure of the divergence from one probability distribution to another. In this paper we seek the divergence from the predicted consideration probabilities to those that are observed in the validation data, recognizing the discrete nature of the data (label = consider or not). For respondent *i* we predict that profile *k* is considered with probability, $r_{ik}=\Pr(y_{ik}=1|\vec{x}_{ik},model)$. Then the divergence from the true model (the y_{ik}'s) to the model being tested (the r_{ik}'s) is given by Equation A1. With log-based-2, KL has the units of bits.

(A1) $$KL = \sum_{k \in validation} \left[y_{ik} \log_2 \left(\frac{y_{ik}}{r_{ik}} \right) + (1 - y_{ik}) \log_2 \left(\frac{1 - y_{ik}}{1 - r_{ik}} \right) \right]$$

When the r_{ik}'s are probabilities, we use them directly in Equation A1. When the r_{ik}'s are themselves discrete we use the observations of false positive and false negative predictions to separate the summation into four components. Let V = the number of profiles in the validation sample, $\hat{C}_v$ = the number of considered validation profiles, F_p = the false positive predictions, and F_n = the false negative predictions. Then the KL divergence is given by the following equation where $S_{c,c}$ is the set of profiles that are considered in the calibration data and considered in the validation data. The sets $S_{c,nc}$, $S_{nc,c}$, and $S_{nc,nc}$ are defined similarly ($nc \rightarrow$ not considered).

$$KL = \sum_{S_{c,c}} \log_2 \left(\frac{\hat{C}_v}{\hat{C}_v - F_p} \right) + \sum_{S_{c,nc}} \log_2 \left(\frac{V - \hat{C}_v}{F_n} \right) + \sum_{S_{nc,c}} \log_2 \left(\frac{\hat{C}_v}{F_p} \right) + \sum_{S_{c,c}} \log_2 \left(\frac{V - \hat{C}_v}{V - \hat{C}_v - F_n} \right)$$

After algebraic simplification, KL divergence can be written as:

(A2) $$KL = \hat{C}_v \log_2 \hat{C}_v + (V - \hat{C}_v) \log_2 (V - \hat{C}_v) - (\hat{C}_v - F_p) \log_2 (\hat{C}_v - F_p) - F_n \log_2 F_n - F_p \log_2 F_p - (V - \hat{C}_v - F_n) \log_2 (V - \hat{C}_v - F_n)$$

KL divergence is a sum over the set of profiles. Sets with more profiles are harder to fit; if V were twice as large and $\hat{C}_v$, F_p, and F_n were scaled proportionally, then the KL divergence would be twice as large. Thus, for comparability across respondents with different validation-set sizes, we scale the KL divergence measure by dividing by V.

Appendix 4. Percent Uncertainty Explained (U^2) for Other Querying Methods