MARKETING SCIENCE
CELEBRATING 30 YEARS
Vol. 30, No. 5, September–October 2011, pp. 801–819
ISSN 0732-2399 | EISSN 1526-548X | 11 | 3005 | 0801

informs®
http://dx.doi.org/10.1287/mksc.1110.0660
© 2011 INFORMS

Active Machine Learning for Consideration Heuristics

Daria Dzyabura, John R. Hauser
MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142
{dariasil@mit.edu, hauser@mit.edu}

We develop and test an active-machine-learning method to select questions adaptively when consumers use heuristic decision rules. The method tailors priors to each consumer based on a "configurator." Subsequent questions maximize information about the decision heuristics (minimize expected posterior entropy). To update posteriors after each question, we approximate the posterior with a variational distribution and use belief propagation (iterative loops of Bayes updating). The method runs sufficiently fast to select new queries in under a second and provides significantly and substantially more information per question than existing methods based on random, market-based, or orthogonal-design questions.

Synthetic data experiments demonstrate that adaptive questions provide close-to-optimal information and outperform existing methods even when there are response errors or "bad" priors. The basic algorithm focuses on conjunctive or disjunctive rules, but we demonstrate generalizations to more complex heuristics and to the use of previous-respondent data to improve consumer-specific priors. We illustrate the algorithm empirically in a Web-based survey conducted by an American automotive manufacturer to study vehicle consideration (872 respondents, 53 feature levels). Adaptive questions outperform market-based questions when estimating heuristic decision rules. Heuristic decision rules predict validation decisions better than compensatory rules.

Key words: active learning; adaptive questions; belief propagation; conjunctive models; consideration sets; consumer heuristics; decision heuristics; disjunctions of conjunctions; lexicographic models; variational Bayes estimation

History: Received: January 6, 2010; accepted: May 4, 2011; Eric Bradlow and then Preyas Desai served as the editor-in-chief and Fred Feinberg served as associate editor for this article. Published online in *Articles in Advance* July 15, 2011.

1. Problem Statement: Adaptive Questions to Identify Heuristic Decision Rules

We develop and test an active-machine-learning algorithm to identify heuristic decision rules. Specifically, we select questions adaptively based on prior beliefs and respondents' answers to previous questions. To the best of our knowledge, this is the first (near-optimal) adaptive-question method focused on consumers' noncompensatory decision heuristics. Extant adaptive methods focus on compensatory decision rules and are unlikely to explore the space of noncompensatory decision rules efficiently (e.g., Evgeniou et al. 2005; Toubia et al. 2007, 2004; Sawtooth 1996). In prior noncompensatory applications, question selection was almost always based on either random profiles or profiles chosen from an orthogonal design.

We focus on noncompensatory heuristics because of managerial and scientific interest. Scientific interest is well established. Experimental and revealed-decision-rule studies suggest that noncompensatory heuristics are common, if not dominant, when consumers face decisions involving many alternatives, many features, or if they are making consideration rather than purchase decisions (e.g., Gigerenzer and Goldstein 1996; Payne et al. 1988, 1993; Yee et al. 2007). Heuristic rules often represent a rational tradeoff among decision costs and benefits and may be more robust under typical decision environments (e.g., Gigerenzer and Todd 1999). Managerial interest is growing as more firms focus product development and marketing efforts on getting consumers to consider their products or, equivalently, preventing consumers from rejecting products without evaluation. We provide illustrative examples in this paper, but published managerial examples include Japanese banks, global positioning systems, desktop computers, smart phones, and cellular phones (Ding et al. 2011, Liberali et al. 2011).

Our focus is on adaptive question selection, but to select questions adaptively, we need intermediate estimates after each answer and before the next question is asked. To avoid excessive delays in online questionnaires, intermediate estimates must be obtained in a second or less (e.g., Toubia et al. 2004). This is a difficult challenge when optimizing questions for noncompensatory heuristics because we must search over a discrete space of the order of 2^N decision rules, where N is the number of feature levels (called aspects, as in Tversky 1972). Without special structure, finding a best-fitting heuristic is much more difficult than finding best-fitting parameters for an (additive) compensatory model—such estimation algorithms typically require the order of N parameters. The ability to scale to large N is important in practice because consideration heuristics are common in product categories with large numbers of aspects (e.g., Payne et al. 1993). Our empirical application searches over 9.0×10^{15} heuristic rules.

We propose an active-machine-learning solution (hereafter, active learning) to select questions adaptively to estimate noncompensatory heuristics. The active-learning algorithm approximates the posterior with a variational distribution and uses belief propagation to update the posterior distribution. It then asks the next question to minimize expected posterior entropy by anticipating the potential responses (in this case, to consider or not consider). The algorithm runs sufficiently fast to be implemented between questions in an online questionnaire.

In the absence of error, this algorithm comes extremely close to the theoretical limit of the information that can be obtained from binary responses. With response errors modeled, the algorithm does substantially and significantly better than extant question-selection methods. We also address looking ahead S steps, generalized heuristics, and the use of population data to improve priors. Synthetic data suggest that the proposed method recovers parameters with fewer questions than extant methods. Empirically, adaptive-question selection is significantly better at predicting future consideration than benchmark question selection. Noncompensatory estimation is also significantly better than the most commonly applied compensatory method.

We begin with a brief review and taxonomy of existing methods to select questions to identify consumer decision rules. We then review noncompensatory heuristics and motivate their managerial importance. Next, we present the algorithm, test parameter recovery with synthetic data, and describe an empirical illustration in the automobile market. We close with generalizations and managerial implications.

2. Existing Methods for Question Selection to Reveal Consumer Decision Rules

Marketing has a long tradition of methods to measure consumer decision rules. Figure 1 attempts a taxonomy that highlights the major trends and provides examples.

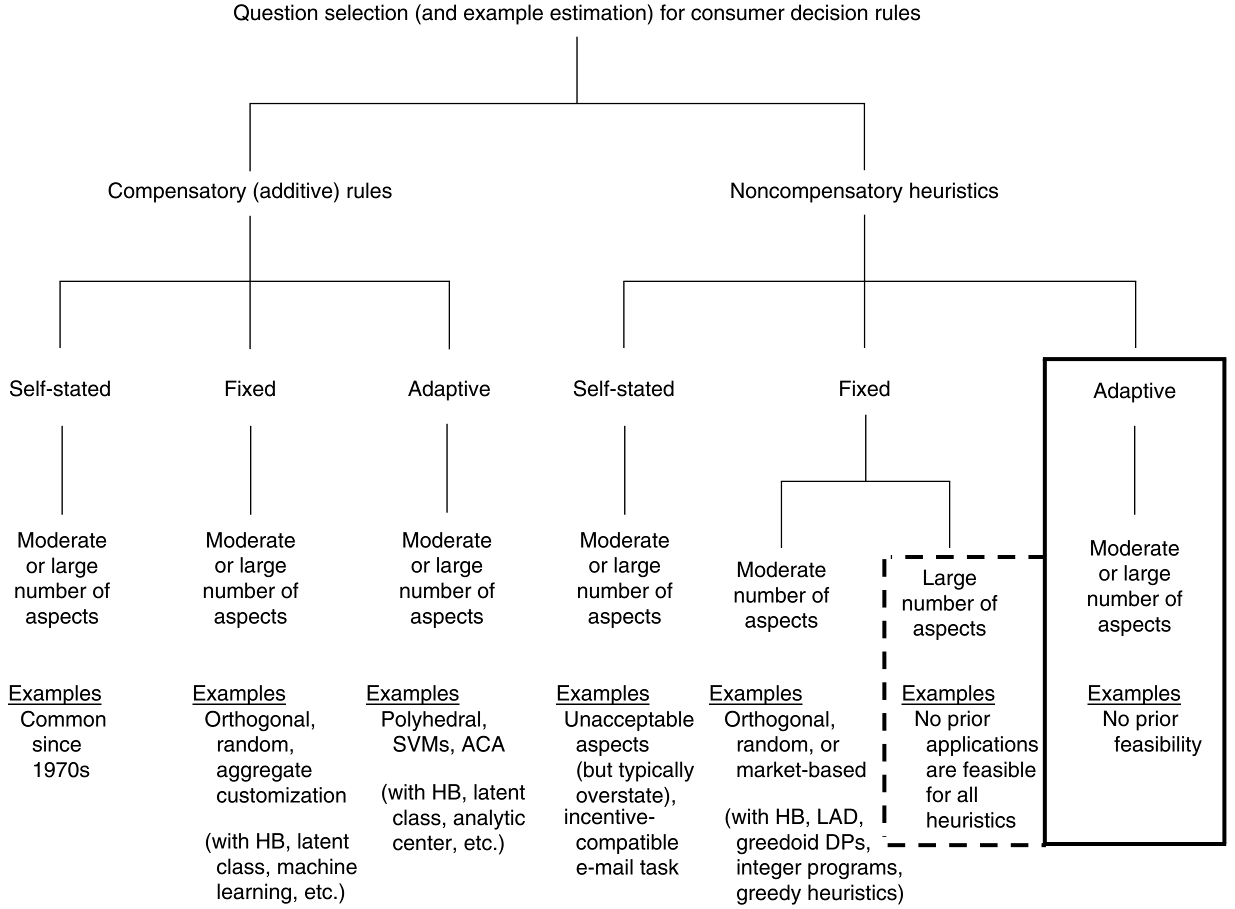
The vast majority of papers focus on compensatory decision rules. The most common methods include either self-explication, which asks respondents to self-state their decision rules, or conjoint analysis, which infers compensatory decision rules from questions in which respondents choose, rank, or rate bundles of aspects called product profiles. These methods are applied widely and have demonstrated both predictive accuracy and managerial relevance (e.g., Green 1984, Green and Srinivasan 1990, Wilkie and Pessemier 1973). In early applications, profiles were chosen from either full-factorial, fractional-factorial, or orthogonal designs, but as hierarchical-Bayes estimation became popular, many researchers moved to random designs to explore interactions better. For choice-based conjoint analysis, efficient designs are a function of the parameters of compensatory decision rules, and researchers developed "aggregate customization" to preselect questions using data from prestudies (e.g., Arora and Huber 2001). More recently, faced with impatient online respondents, researchers developed algorithms for adaptive conjoint questions based on compensatory models (e.g., Toubia et al. 2004). After data are collected adaptively, the likelihood principle enables the data to be reanalyzed with models using classical statistics, Bayesian statistics, or machine learning.

In some applications respondents are asked to self-state noncompensatory heuristics. Self-explication has had mixed success because respondents often chose profiles with aspects they had previously stated as unacceptable (e.g., Green et al. 1988). Recent experiments with incentive-compatible tasks, such as having respondents write an e-mail to a friend who will act as their agent, are promising (Ding et al. 2011).

Researchers have begun to propose methods to identify heuristic decision rules from directly measured consideration of product profiles. Finding the best-fit decision rule requires solving a discrete optimization problem that is NP-hard (e.g., Martignon and Hoffrage 2002). Existing estimation uses machine-learning methods such as greedy heuristics, greedoid dynamic programs, logical analysis of data, or linear programming perturbation (Dieckmann et al. 2009, Hauser et al. 2010, Kohli and Jedidi 2007, Yee et al. 2007). Even for approximate solutions, runtimes are exponential in the number of aspects limiting methods to moderate numbers of aspects. Bayesian methods have been used to estimate parameters for moderate numbers of aspects (e.g., Gilbride and Allenby 2004, 2006; Hauser et al. 2010; Liu and Arora 2011). To date, profiles for direct consideration measures are chosen randomly or from an orthogonal design.

Within this taxonomy Figure 1 illustrates the focus of this paper (thick box)—adaptive questions for noncompensatory heuristics. We also develop an estimation method for noncompensatory heuristics that scales to large numbers of aspects, even when applied

Figure 1 Taxonomy of Existing Methods to Select Questions to Identify Consumer Decision Rules


Question selection (and example estimation) for consumer decision rules
Compensatory (additive) rules
Noncompensatory heuristics
Self-stated
Fixed
Adaptive
Self-stated
Fixed
Adaptive
Moderate or large number of aspects
Moderate or large number of aspects
Moderate or large number of aspects
Moderate or large number of aspects
Moderate number of aspects
Large number of aspects
Moderate or large number of aspects
Examples
Common since 1970s
Examples
Orthogonal, random, aggregate customization
(with HB, latent class, machine learning, etc.)
Examples
Polyhedral, SVMs, ACA
(with HB, latent class, analytic center, etc.)
Examples
Unacceptable aspects (but typically overstate), incentive-compatible e-mail task
Examples
Orthogonal, random, or market-based
(with HB, LAD, greedoid DPs, integer programs, greedy heuristics)
Examples
No prior applications are feasible for all heuristics
Examples
No prior feasibility

Note. ACA, adaptive conjoint analysis; DP, dynamic program; HB, hierarchical Bayes; SVMs, support-vector machines.

to extant question-selection methods (dotted box). We focus on questions that ask about consideration directly (consider or not). However, our methods apply to all data in which the consumer responds with a yes or no answer and might be extendable to choice-based data where more than one profile is shown at a time. We do not focus on methods where consideration is an unobserved construct inferred from choice data (e.g., Erdem and Swait 2004, van Nierop et al. 2010). There is one related adaptive method—the first stage of adaptive choice-based conjoint analysis (ACBC; Sawtooth 2008) that is based on rules of thumb to select approximately 28 profiles that are variations on a "bring-your-own" profile. Profiles are not chosen optimally, and noncompensatory heuristics are not estimated.

3. Noncompensatory Decision Heuristics

We classify decision heuristics as simple and more complex. The simple heuristics include conjunctive, disjunctive, lexicographic, take-the-best, and elimination by aspects. The more complex heuristics include subset conjunctive and disjunctions of conjunctions. The vast majority of scientific experiments have examined the simple heuristics, with conjunctive the most common (e.g., Gigerenzer and Selten 1999; Payne et al. 1988, 1993, and references therein). The study of more complex heuristics, which nest the simple heuristics, is relatively recent, but there is evidence that some consumers use the more complex forms (Jedidi and Kohli 2005, Hauser et al. 2010). Both simple and complex heuristics apply for consideration, choice, or other decisions and for a wide variety of product categories. For simplicity of exposition, we define the heuristics with respect to the consideration decision and illustrate the heuristics for automotive features.

3.1. Simple Heuristics

3.1.1. Conjunctive Heuristic. For some features consumers require acceptable ("must-have") levels. For example, a consumer might only consider a sedan body type and only consider Toyota, Nissan, or Honda. Technically, for features not in the conjunction, such as engine type, all levels are acceptable.

3.1.2. Disjunctive Heuristic. If the product has "excitement" levels of a feature, the product is considered no matter what the levels of the other features are. For example, a consumer might consider all vehicles with a hybrid engine.

3.1.3. Take-the-Best. The consumer ranks products on a single most diagnostic feature and considers only those above some cutoff. For example, the consumer may find "brand" most diagnostic, rank products on brand, and consider only those with brands that are acceptable—say, Toyota, Nissan, and Honda.

3.1.4. Lexicographic (by Features). This heuristic is similar to take-the-best except the feature need not be most diagnostic. If products are tied on a feature level, then the consumer continues examining features lower in the lexico ordering until ties are broken. For example, the consumer might rank on brand, then body style considering only Toyota, Nissan, and Honda, and, among those brands, only sedans.

3.1.5. Elimination by Aspects. The consumer selects an aspect and eliminates all products with unacceptable levels, and then he or she selects another aspect and eliminates products with unacceptable levels on that aspect, continuing until only considered products are left. For example, the consumer may eliminate all but Toyota, Nissan, and Honda and all but sedans. Researchers have also examined acceptance by aspects and lexicographic by aspects that generalize elimination by aspects in the obvious ways.

When the only data are consider versus not consider, it does not matter in which order the profiles were eliminated or accepted. Take-the-best, lexicographic (by features), elimination by aspects, acceptance by aspects, and lexicographic by aspects are indistinguishable from conjunctive heuristics. The rules predict differently when respondents are asked to rank data and differ in the underlying cognitive process, but they do not differ when predicting the observed consideration set. Disjunctive is a mirror image of conjunctive. Thus, any question-selection algorithm that optimizes questions to identify conjunctive heuristics can be applied (perhaps with a mirror image) to any of the simple heuristics.

3.2. More Complex Heuristics

3.2.1. Subset Conjunctive. The consumer considers a product if F features have levels that are acceptable. The consumer does not require all features to have acceptable levels. For example, the consumer might have acceptable brands (Toyota, Honda, Nissan), acceptable body types (sedan), and acceptable engines (hybrid) but only require that two of the three features have levels that are acceptable.

3.2.2. Disjunctions of Conjunctions. The consumer might have two or more sets of acceptable aspects. For example, the consumer might consider [Toyota and Honda sedans] or [crossover body types with hybrid engines]. Disjunctions of conjunctions nests the subset conjunctive heuristic and all of the simple heuristics (for consideration). However, its generality is also a curse. Empirical applications require cognitive simplicity to avoid overfitting data.

All of these decision heuristics are postulated as descriptions of how consumers make decisions. Heuristics are not, and need not be, tied to utility maximization. For example, it is perfectly reasonable for a consumer to screen out low-priced products because the consumer believes that he or she is unlikely to choose such a product if considered and, hence, does not believe that evaluating such a product is worth the time and effort. (Put another way, the consumer would purchase a fantastic product at a low price if he or she knew about the product but never finds out about the product because the consumer chose not to evaluate low-priced products. When search costs are considered, it may be rational for the consumer not to search the lower-priced product because the probability of finding an acceptable low-priced product is too low.)

In this paper we illustrate our question-selection algorithm with conjunctive decision rules (hence it applies to all simple heuristics). We later extend the algorithm to identify disjunctions-of-conjunctions heuristics (which nest subset conjunctive heuristics).

4. Managerial Relevance: Stylized Motivating Example

As a stylized example, suppose that automobiles can be described by four features with two levels each: Toyota or Chevy, sedan or crossover body type, hybrid or gasoline engine, and premium or basic trim levels, for a total of eight aspects. Suppose we are managing the Chevy brand that makes only sedans with gasoline engines and basic trim, and suppose it is easy to change trim levels but not the other features. If consumers are compensatory and their partworths are heterogeneous and not "too extreme," we can get some consumers to consider our vehicle by offering sufficiently premium trim levels. It might be profitable to do so.

Suppose instead that a segment of consumers is conjunctive on [Toyota $\wedge$ crossover]. (In our notation, $\wedge$ is the logical "and"; $\vee$ is the logical "or.") No amount of trim levels will attract these conjunctive consumers. They will not pay attention to Chevy advertising, visit the GM website, or travel to a Chevy dealer—they will never evaluate any Chevy sedans no matter how much we improve them. In another example, if a segment of consumers is conjunctive on [crossover ∧ hybrid], we will never get those consumers to evaluate our vehicles unless we offer a hybrid crossover vehicle no matter how good we make our gasoline-engine sedan. Even with disjunctions of conjunctions, consumers who use [(sedan ∧ hybrid) ∨ (crossover ∧ gasoline engine)] will never consider our gasoline-engine sedan. In theory we might approximate noncompensatory heuristics with compensatory partworth decision rules (especially if we include interactions), but if there are many aspects, empirical approximations may not be accurate.

Many products just never make it because they are never considered; consumers never learn that the products have outstanding aspects that could compensate for the product's lack of a conjunctive feature. Our empirical illustration is based in the automotive industry. Managers at high levels in the sponsoring organization believe that conjunctive screening was a major reason that the automotive manufacturer faced slow sales relative to other manufacturers. For example, they had evidence that more than half of the consumers in the United States would not even consider their brands. Estimates of noncompensatory heuristics are now important inputs to product-design and marketing decisions at that automotive manufacturer.

Noncompensatory heuristics can imply different managerial decisions. Hauser et al. (2010) illustrate how rebranding can improve the share of a common electronic device if consumers use compensatory models but not if consumers use noncompensatory models. Ding et al. (2011) illustrate that conjunctive rules and compensatory rules are correlated in the sense that feature levels with higher average partworth values also appear in more "must-have rules." However, the noncompensatory models identify combinations of aspects that would not be considered even though their combined partworth values might be reasonable.

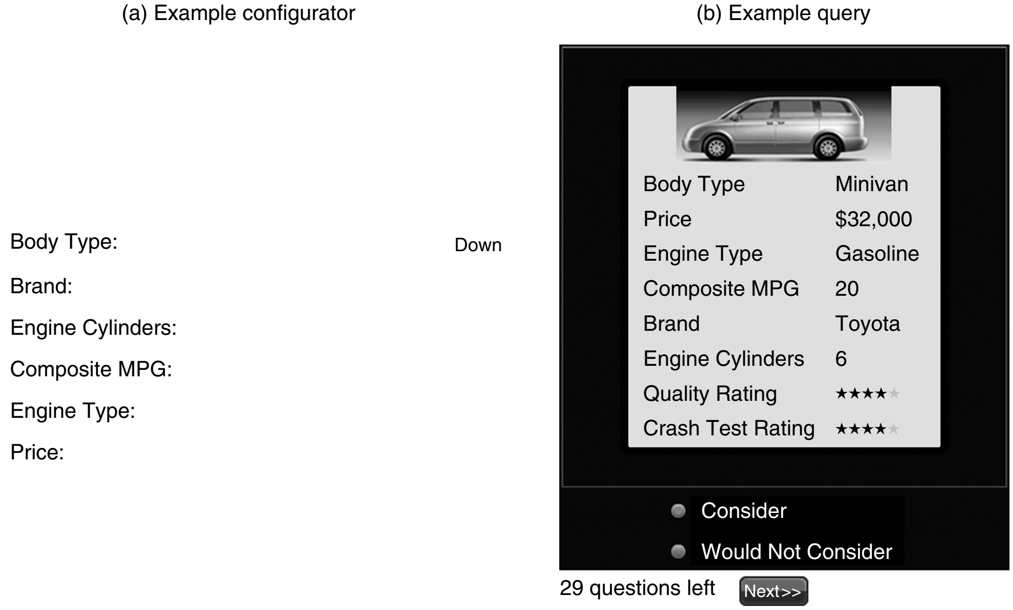
5. Question Types, Error Structure, and Notation

5.1. Question Types and Illustrative Example

Figure 2 illustrates the basic question formats. The example is automobiles, but these types of questions have been used in a variety of product categories—usually durable goods, where consideration is easy to define and a salient concept to consumers (Dahan and Hauser 2002, Sawtooth 2008, Urban and Hauser 2004). Extensive pretests suggest that respondents can accurately "configure" a profile that they would consider (Figure 2(a)). If respondents use only one conjunctive rule in their heuristic, they find it difficult to accurately configure a second profile. If they use a disjunctions-of-conjunctions heuristic with sufficiently distinct conjunctions, such as [(Toyota ∧ sedan) ∨ (Chevy ∧ truck)], we believe they can configure a second profile "that is different from previous profiles that you said you will consider." In this section we focus on the first configured profile and the corresponding conjunctive heuristic. In a later section, we address more conjunctions in a disjunctions-of-conjunctions heuristic.

After configuring a considered profile, we ask respondents whether or not they will consider various

Figure 2 **Example Configurator and Example Queries (Color in Original)**



(a) Example configurator

(b) Example query

Body Type	Minivan
Price	$32,000
Engine Type	Gasoline
Composite MPG	20
Brand	Toyota
Engine Cylinders	6
Quality Rating	★★★★☆
Crash Test Rating	★★★★☆

profiles (Figure 2(b)). Our goal is to select the profiles that provide the most information about decision heuristics (information is defined below). With synthetic data we plot cumulative information (parameter recovery) as a function of the number of questions. In our empirical test, we ask 29 queries, half of which are adaptive and half of which are chosen randomly (proportional to market share). We compare predictions based on the two types of questions. Although the number of questions was fixed in the empirical test, we address how stopping rules can be endogenous to the algorithm.

5.2. Notation, Error Structure, and Question-Selection Goal

Let M be the number of features (e.g., brand, body style, engine type, trim level; $M = 4$), and let N be the total numbers of aspects (e.g., Toyota, Chevy, sedan, crossover, hybrid, gasoline engine, low trim, high trim; $N = 8$). Let i index consumers and j index aspects. For each conjunction, 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. For example, $\vec{a}_i = \{+1, -1, +1, -1, +1, -1, +1, +1\}$ would indicate that the ith consumer finds hybrid Toyota sedans with both low and high trim to be acceptable.

Each sequential query (Figure 2(b)), indexed by k, is a profile, $\vec{x}_{ik}$, with N elements, x_{ijk}, such that $x_{ijk} = 1$ if i's profile k has aspect j and $x_{ijk} = 0$ if it does not. Each $\vec{x}_{ik}$ has exactly M nonzero elements, one for each feature. (In our stylized example, a profile contains one brand, one body type, one engine type, and one trim level.) For example, $\vec{x}_{ik} = \{1, 0, 1, 0, 1, 0, 1, 0\}$ would be a hybrid Toyota sedan with low trim.

Let X_{iK} be the matrix of the first K profiles given to a consumer; each row corresponds to a profile. Mathematically, profile $\vec{x}_{ik}$ *satisfies* a conjunctive rule $\vec{a}_i$ if whenever $x_{ijk} = 1$, then $a_{ij} = 1$, such that every aspect of the profile is acceptable. In our eight-aspect example, consumer i finds the hybrid Toyota sedan with low trim to be acceptable (compare $\vec{a}_i$ to $\vec{x}_{ik}$). This condition can be expressed as $\min_j\{x_{ijk}a_{ij}\} \geq 0$. It is violated only if a profile has at least one level ($x_{ijk} = 1$) that is unacceptable ($a_{ij} = -1$). Following Gilbride and Allenby (2004), we define a function to indicate when a profile is acceptable: $I(\vec{x}_{ik}, \vec{a}_i) = 1$ if $\min_j\{x_{ijk}a_{ij}\} \geq 0$, and $I(\vec{x}_{ik}, \vec{a}_i) = 0$ otherwise. We use the same coding for disjunctive rules but modify the definition of $I(\vec{x}_{ik}, \vec{a}_i)$ to use $\max_j$ rather than $\min_j$.

Let y_{ik} be consumer i's answer to the kth query, 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 answers. If there were no response errors, we would observe $y_{ik} = 1$ if and only if $I(\vec{x}_{ik}, \vec{a}_i) = 1$. However, empirically, we expect response errors. Because the algorithm must run rapidly between queries, we choose a simple form for response error. Specifically, we assume that a consumer gives a false-positive answer with probability ϵ_1 and a false-negative answer with probability ϵ_2. For example, the ith consumer will say "consider ($y_{ik} = 1$)" with probability $1 - \epsilon_2$ whenever the indicator function implies "consider," but he or she will also say "consider" with probability ϵ_1 if the indicator function implies "not consider." This error structure implies the following data-generating model:

$$\begin{aligned} \Pr(y_{ik}=1 \mid \vec{x}_{ik}, \vec{a}_i) &= (1-\epsilon_2)I(\vec{x}_{ik}, \vec{a}_i) + \epsilon_1(1-I(\vec{x}_{ik}, \vec{a}_i)), \\ \Pr(y_{ik}=0 \mid \vec{x}_{ik}, \vec{a}_i) &= \epsilon_2 I(\vec{x}_{ik}, \vec{a}_i) + (1-\epsilon_1)(1-I(\vec{x}_{ik}, \vec{a}_i)). \end{aligned} \tag{1}$$

Each new query $\vec{x}_{i,K+1}$ is based on our posterior beliefs about the decision rules ($\vec{a}_i$). After the Kth query, we compute the posterior $\Pr(\vec{a}_i \mid X_{iK}, y_{iK})$ conditioned on the first K queries (X_{iK}), the first K answers ($\vec{y}_{iK}$), and the priors. (Posterior beliefs might also reflect information from other respondents; see §6.6.) We seek to select the $\vec{x}_{ik}$s to get as much information as feasible about $\vec{a}_i$ or, equivalently, to reduce uncertainty about $\vec{a}_i$ by the greatest amount. In §6.2 we define "information" and describe how we optimize it.

5.3. Error Magnitudes Are Set Prior to Data Collection

We cannot know the error magnitudes until after data are collected, but we must set the ϵs in order to collect data. Setting the ϵs is analogous to setting "accuracy" parameters in aggregate customization. We address this conundrum in two ways: (1) We treat the ϵs as "tuning" parameters and explore the sensitivity to these tuning parameters with synthetic data. Setting tuning parameters is common in machine-learning query selection. (2) For the empirical test, we rely on managerial judgment (Little 2004a, b). Because the tuning parameters are set by managerial judgment prior to data collection, our empirical test is conservative in the sense that predictions might improve if future research allows updating of the error magnitudes within or across respondents.

To aid intuition we motivate the ϵs with an illustrative microanalysis of our stylized example. Suppose that a respondent's conjunctive heuristic is [Toyota $\wedge$ crossover]. This respondent should find a crossover Toyota acceptable and not care about the engine and trim. Coding each aspect as acceptable or not, and preserving the order Toyota, Chevy, sedan, crossover, hybrid, gasoline, premium trim, and basic trim, this heuristic becomes $\vec{a}_i = [+1, -1, -1, +1, +1, +1, +1, +1]$. Suppose that when this respondent makes a consideration decision, he or she makes errors with probability η on each aspect, where an error involves flipping that aspect's acceptability. For example, suppose he or she is shown a Toyota crossover with a hybrid engine and premium trim; that is, $\vec{x}_{ik} = [1, 0, 0, 1, 1, 0, 1, 0]$. He or she matches the heuristic to the profile aspect by aspect, making errors with probability η for each acceptable aspect in the profile; e.g., Toyota is acceptable per the heuristic but may be mistaken for unacceptable with probability η. The respondent can make a false-negative error if any of the four aspects in the profile are mistaken for unacceptable ones. If these errors occur independently, the respondent will make a false-negative error with probability $\epsilon_2 = 1 - (1 - \eta)^4$. If a profile is unacceptable, say, $\vec{x}_{ik} = [0, 1, 1, 0, 1, 0, 1, 0]$, we easily compute $\epsilon_1 = \eta^2(1-\eta)^2$.

In this illustration, any prior belief on the distribution of the heuristics and profiles implies expected ϵs as a function of the ηs. Whether one specifies the ηs and derives expected ϵs or specifies the ϵs directly depends on the researchers and managers, but in either case, the tuning parameters are specified prior to data collection. With synthetic data we found no indication that one specification is preferred to the other. Empirically, we found it easier to think about the ϵs directly.

6. Adaptive Question Selection

To select questions adaptively, we must address the following procedure:

Step 1. Initialize beliefs by generating consumer-specific priors.

Step 2. Select the next query based on current posterior beliefs.

Step 3. Update posterior beliefs from the priors and the responses to all the previous questions.

Step 4. Continue looping Steps 2 and 3 until Q questions are asked (or until another stopping rule is reached).

6.1. Initialize Consumer-Specific Beliefs (Step 1)

Hauser and Wernerfelt (1990, p. 393) provide examples where self-stated consideration set sizes are one-tenth or less of the number of brands on the market. Our experience suggests these examples are typical. If the question-selection algorithm used noninformative priors, the initial queries would be close to random guesses, most of which would 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, the first considered profile provides substantial information and a significant shift in beliefs. Without observing the first considered profile directly, queries are not efficient, particularly with large numbers of aspects (N). To address this issue, we ask each respondent to configure a considered profile and, hence, gain substantial information.

Prior research using compensatory rules (e.g., Toubia et al. 2004) suggests that adaptive questions are most efficient relative to random or orthogonal questions when consumers' heuristic decision rules are heterogeneous. We expect similar results for noncompensatory heuristics. In the presence of heterogeneity, the initial configured profile enables us to tailor prior beliefs to each respondent.

For example, in our empirical application we tailor prior beliefs using the co-occurrence of brands in consideration sets. Such data are readily available in the automotive industry and for frequently purchased consumer goods. Alternatively, prior beliefs might be updated on the fly using a collaborative filter on prior respondents (see §6.6). Without loss of generality, let $j = 1$ index the brand aspect the respondent configures, and for other brand aspects, let b_{1j} be the prior probability that brand j is acceptable when brand 1 is acceptable. Let $\vec{x}_{i1}$ be the configured profile, and set $y_{i1} = 1$. When co-occurrence data are available, prior beliefs on the marginal probabilities are set such that $\Pr(a_{i1} = 1 \mid \vec{x}_{i1}, y_{i1}) = 1$ and $\Pr(a_{ij} = 1 \mid \vec{x}_{i1}, y_{i1}, priors) = b_{ij}$ for $j \neq 1$.

Even without co-occurrence data, we can set respondent-specific priors for every aspect on which we have strong prior beliefs. We use weakly informative priors for all other aspects. When managers have priors across features (e.g., considered hybrids are more likely to be Toyotas), we also incorporate those priors (Little 2004a, b).

6.2. Select the Next Question Based on Posterior Beliefs from Prior Answers (Step 2)

The respondent's answer to the configurator provides the first of a series of estimates of his or her decision rule, $p_{ij1} = \Pr(a_{ij} = 1 \mid X_{i1} = \vec{x}_{i1}, \vec{y}_{i1})$ for all aspects j. (We have suppressed the notation for "priors.") We update these probabilities by iterating through Steps 2 and 3, computing updated estimates after each question–answer pair using all data collected up to and including that the Kth question, $p_{ijK} = \Pr(a_{ij} = 1 \mid X_{iK}, \vec{y}_{iK})$ for $K > 1$. (Details are in Step 3; see §6.3.) To select the $K+1$st query (Step 2), assume we have computed posterior values (p_{ijK}) from prior queries (up to K) and that we can compute contingent values ($p_{ij,\,K+1}$) one step ahead for any potential new query ($\vec{x}_{i,\,K+1}$) and its corresponding answer ($y_{i,\,K+1}$). We seek those questions that tell us as much as feasible about the respondent's decision heuristic. Equivalently, we seek to reduce uncertainty about $\vec{a}_i$ by the greatest amount.

Following Lindley (1956) we define the most informative question as the query that minimizes a loss function. In this paper, we use Shannon's entropy as the uncertainty measure (Shannon 1948), but other measures of uncertainty could be used without otherwise changing the algorithm. Shannon's entropy, measured in *bits*, quantifies the amount of information that is missing because the value of a random variable is not known for certain. Higher entropy corresponds to more uncertainty. Zero entropy corresponds to perfect knowledge. Shannon's entropy (hereafter, entropy) is used widely in machine learning, has proven robust in many situations, and is the basis of criteria used to evaluate parameter recovery and predictive ability (U^2 and Kullback–Leibler 1951 divergence). We leave to future implementations other loss functions such as Rényi (1961) entropy, suprisals, and other measures of information.[1] Mathematically,

$$H_{\vec{a}_i} = \sum_{j=1}^{N} -\{p_{ijK}\log_2 p_{ijK} + (1-p_{ijK})\log_2(1-p_{ijK})\}. \quad (2)$$

If some aspects are more important to managerial strategy, we use a weighted sum in Equation (2).

To select the $K+1$st query, $\vec{x}_{i,K+1}$, we enumerate candidate queries, anticipating the answer to the question, $y_{i,K+1}$, and anticipating how that answer updates our posterior beliefs about the respondent's heuristic. Using the p_{ijK}s we compute the probability the respondent will consider the profile, $q_{i,K+1}(\vec{x}_{i,K+1}) = \Pr(y_{i,K+1}=1 \mid X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1})$. Using the Step 3 algorithm (described in the next subsection), we update the posterior $p_{ij,K+1}$s for all potential queries and answers. Let $p^{+}_{ij,K+1}(\vec{x}_{i,K+1}) = \Pr(a_{ij}=1 \mid X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1}, y_{i,K+1}=1)$ be the posterior beliefs if we ask profile $\vec{x}_{i,K+1}$ and the respondent considers it. Let $p^{-}_{ij,K+1}(\vec{x}_{i,K+1}) = \Pr(a_{ij}=1 \mid X_{iK}, \vec{y}_{iK}, \vec{x}_{i,K+1}, y_{i,K+1}=-1)$ be the posterior beliefs if the respondent does not consider the profile. Then the expected posterior entropy is

$$\begin{aligned} &E[H_{\vec{a}i}(\vec{x}_{i,K+1} \mid X_{iK}, \vec{Y}_{iK})] \\ &= -q_{i,K+1}(\vec{x}_{i,K+1})\sum_j \{p^{+}_{ij,K+1}(\vec{x}_{i,K+1})\log_2[p^{+}_{ij,K+1}(\vec{x}_{i,K+1})] \\ &\qquad + [1-p^{+}_{ij,K+1}(\vec{x}_{i,K+1})]\log_2[1-p^{+}_{ij,K+1}(\vec{x}_{i,K+1})]\} \\ &\quad -[1-q_{i,K+1}(\vec{x}_{i,K+1})]\sum_j \{p^{-}_{ij,K+1}(\vec{x}_{i,K+1})\log_2[p^{-}_{ij,K+1}(\vec{x}_{i,K+1})] \\ &\qquad + [1-p^{-}_{ij,K+1}(\vec{x}_{i,K+1})]\log_2[1-p^{-}_{ij,K+1}(\vec{x}_{i,K+1})]\}. \quad (3) \end{aligned}$$

When the number of feasible profiles is moderate, we compute Equation (3) for every profile and choose the profile that minimizes Equation (3). However, in large designs such as the 53-aspect design in our empirical example, the number of potential queries (357,210) can be quite large. Because this large number of computations cannot be completed in less than a second, we focus our search using *uncertainty sampling* (e.g., Lewis and Gale 1994). Specifically, we evaluate Equation (3) for the T queries about which we are most uncertain. "Most uncertain" is defined as $q_{i,K+1}(\vec{x}_{i,K+1}) \approx 0.5$. Profiles identified from among the T most uncertain profiles are approximately optimal and, in some cases, optimal (e.g., see Appendix A). Uncertainty sampling is similar to choice balance as used in both polyhedral methods and aggregate customization (e.g., Arora and Huber 2001, Toubia et al. 2004). Synthetic data tests demonstrate that with T sufficiently large, we achieve close-to-optimal expected posterior entropy. For our empirical application, setting $T = 1{,}000$ kept question selection under a second. As computing speeds improve, researchers can use a larger T.

Equation (3) is myopic because it computes expected posterior entropy one step ahead. Extending the algorithm S steps ahead is feasible for small N. However S-step computations are exponential in S. For example, if there are 256 potential queries, a two-step ahead algorithm requires that we evaluate $256^2 = 65{,}536$ potential queries (without further approximations). Fortunately, synthetic data experiments suggest that one-step ahead computations achieve close to the theoretical maximum information of one bit per query (when there are no response errors) and do quite well when there are response errors. For completeness we coded a two-step-ahead algorithm in the case of 256 potential queries. Even for modest problems, its running time was excessive (over 13 minutes between questions); it provided negligible improvements in parameter recovery. Our empirical application has over a thousand times as many potential queries—a two-step-ahead algorithm was not feasible computationally.

6.3. Update Beliefs About Heuristic Rules Based on Answers to the K Questions (Step 3)

In Step 3 we use Bayes theorem to update our beliefs after the Kth query:

$$\begin{aligned} &\Pr(\vec{a}_i \mid X_{iK}, \vec{y}_{iK}) \\ &\quad \propto \Pr(y_{iK} \mid \vec{x}_{iK}, \vec{a}_i = \vec{a})\Pr(\vec{a} = \vec{a}_i \mid X_{i,K-1}, \vec{y}_{i,K-1}). \quad (4) \end{aligned}$$

The likelihood term, $\Pr(y_{iK} \mid \vec{x}_{iK}, \vec{a}_i = \vec{a})$, comes from the data-generating model in Equation (1). The variable of interest, $\vec{a}_i$, is defined over all binary vectors of length N. Because the number of potential conjunctions is exponential in N, updating is not computationally feasible without further structure on the distribution of conjunctions. For example, with

[1] Rényi's entropy reduces to Shannon's entropy when Rényi's $\alpha = 1$; the only value of α for which information on the a_{ij}s is separable. To use these measures of entropy, modify Equations (2) and (3) to reflect Rényi's α.

$N = 53$ in our empirical example, we would need to update the distribution for 9.0×10^{15} potential conjunctions.

To gain insight for a feasible algorithm, we examine solutions to related problems. Gilbride and Allenby (2004) use a "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. Even for vertical features, such as price, we want to allow non-threshold heuristics. We need algorithms that let us classify each level as acceptable or not.

For a feasible algorithm, we use a variational Bayes approach. In variational Bayes inference, a complex posterior distribution is approximated with a variational distribution chosen from a family of distributions judged similar to the true posterior distribution. Ideally, the variational family can be evaluated quickly (Attias 1999, Ghahramani and Beal 2000). Even with an uncertainty-sampling approximation in Step 2, we must compute posterior distributions for $2T$ question–answer combinations and do so while the respondent waits for the next question.

As our variational distribution, we approximate the distribution of $\vec{a}_i$ with N independent binomial distributions. This variational distribution has N parameters, the p_{ij}s, rather than parameters for the 2^N potential values of $\vec{a}_i$. 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. For example, we might find that those automotive consumers who screen on Toyota also screen on hybrid engines. In another application we might find that those cellular phone consumers who screen on Nokia also screen on "flip." For every respondent the posterior values of all p_{ijK}s depend on all of the data from that respondent, not just queries that involve the jth aspect.

To calculate posteriors for the variational distribution, we use a version of belief propagation (Yedidia et al. 2003, Ghahramani and Beal 2001). The algorithm converges iteratively to an estimate of $\vec{p}_{iK}$. The hth iteration uses Bayes theorem to update each p^h_{ijK} based on the data and based on $p^h_{ij'K}$ for all $j' \neq j$. Within the hth iteration, the algorithm loops over aspects and queries using the data-generating model (Equation (1)) to compute the likelihood of observing $y_k = 1$ conditioned on the likelihood for $k' \neq k$. It continues until the estimates of the p^h_{ijK}s stabilize. In our experience, the algorithm converges quickly: 95.6% of the estimations converge in 20 or fewer iterations, 99.2% in 40 or fewer iterations, and 99.7% in 60 or fewer iterations. Appendix B provides the pseudo-code.

Although variational distributions work well in a variety of applications, there is no guarantee for our application. Performance is an empirical question that we address in §§7 and 8. Finally, we note that the belief propagation algorithm and Equation (4) appear to be explicitly dependent only on the questions that are answered by consumer i. However, our notation has suppressed the dependence on prior beliefs. It is a simple matter to make prior beliefs dependent on the distribution of the $\vec{a}_i$s, as estimated from previous respondents (see §6.6).

6.4. Stopping Rules (Step 4)

Adaptive-question selection algorithms for compensatory decision rules and fixed question-selection algorithms for compensatory or noncompensatory rules rely on a target number of questions chosen by prior experience or judgment. Such a stopping rule can be used with the adaptive question-selection algorithm proposed in this paper. For example, we stopped after $Q = 29$ questions in our empirical illustration.

However, expected posterior entropy minimization makes it feasible to select a stopping rule endogenously. One possibility is to stop questioning when the expected reduction in entropy drops below a threshold for two or more adaptive questions. Synthetic data provide some insight. In §7 we plot the information obtained about parameters as a function of the number of questions. In theory we might also gain insight from our empirical example. However, because our empirical example used only 29 questions for 53 aspects, for 99% of the respondents the adaptive-question selection algorithm would still have gained substantial information if the respondents had been asked a 30th question. We return to this issue in §11. Because we cannot redo our empirical example, we leave this and other stopping-rule extensions to future research.

6.5. Extension to Disjunctions of Conjunctions

Disjunctions-of-conjunctions heuristics nest both simple and complex heuristics. The extension to disjunctions of conjunctions is conceptually simple. After we reach a stopping rule, whether it be fixed a priori or endogenous, we simply restart the algorithm by asking a second configurator question but requiring an answer that is substantially different from the profiles that the respondent has already indicated he or she will consider. If the respondent cannot configure such a profile, we stop. Empirically, cognitive simplicity suggests that respondents use relatively few conjunctions (e.g., Gigerenzer and Goldstein 1996, Hauser et al. 2010, Martignon and Hoffrage 2002). Most consumers use one conjunction (Hauser et al. 2010).

Hence the number of questions should remain within reason. We test this procedure on synthetic data and, to the extent that our data allow, empirically.

6.6. Using Data from Previous Respondents

We can use data from other respondents to improve priors for new respondents, but in doing so, we want to retain the advantage of consumer-specific priors. Collaborative filtering provides a feasible method (e.g., Breese et al. 1998). We base our collaborative filter on the consumer-specific data available from the configurator (Figure 2(a)).

Specifically, after a new respondent completes the configurator, we use collaboratively filtered data from previous respondents who configured similar profiles. For example, if an automotive consumer configures a Chevy, we search for previous respondents who configured a Chevy. For other brands we compute priors with a weighted average of the brand posteriors (p_{ij}s) from those respondents. (We weigh previous respondents by predictive precision.) We do this for all configured features. As sample sizes increase, population data overwhelm even "bad" priors; performance will converge to performance based on accurate priors (assuming the collaborative filter is effective). We test finite-sample properties on synthetic data and, empirically, with an approximation based on the data we collected.

7. Synthetic Data Experiments

To evaluate the ability of the active-learning algorithm to recover known heuristic decision rules, we use synthetic respondents. To compare adaptive-question selection to established methods, we choose a synthetic decision task with sufficiently many aspects to challenge the algorithm but for which existing methods are feasible. With four features at four levels (16 aspects), there are 65,536 heuristic rules—a challenging problem for extant *heuristic-rule* estimation methods. An orthogonal design is 32 profiles and, hence, in the range of tasks in the empirical literature. We simulate respondents who answer any number of questions $K \in [1, 256]$, where 256 profiles exhaust the feasible profiles. To evaluate the question-selection methods, we randomly select 1,000 heuristic rules (synthetic respondents). For each aspect we draw a Bernoulli probability from a Beta$(1, 1)$ distribution (uniform distribution) and draw a $+1$ or -1 using the Bernoulli probability. This "sample size" is on the high side of what we might expect in an empirical study and provides sufficient heterogeneity in heuristic rules.

For each decision heuristic, $\vec{a}_i$, we use either the proposed algorithm or an established method to select questions. The synthetic respondent then "answers" the questions using the decision heuristic, but with response errors ϵ_1 and ϵ_2 chosen as if generated by reasonable ηs. To compare question-selection methods, we keep the estimation method constant. We use the variational Bayes belief-propagation method developed in this paper. The benchmark question-selection methods are orthogonal, random, and market based. Market-based questions are chosen randomly but in proportion to profile shares we might expect in the market—market shares are known for synthetic data.

With synthetic data we know the parameters a_{ij}. For any K and for all i and j, we use the "observed" synthetic data to update the probability p_{ijK} that $a_{ij} = 1$. An appropriate information-theoretic measure of parameter recovery is U^2, which quantifies the percentage of uncertainty explained (empirical information/initial entropy; Hauser 1978); $U^2 = 100\%$ indicates perfect parameter recovery.

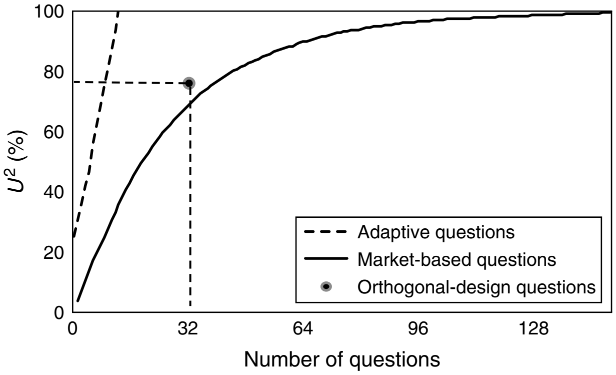
We begin with synthetic data that contain no response errors. These data quantify potential maximum gains with adaptive questions, test how rapidly active-learning questions recover parameters perfectly, and bound improvements that would be possible with nonmyopic S-step-ahead algorithms. We then repeat the experiments with error-laden synthetic data and with "bad" priors. Finally, we examine whether we can recover disjunctions-of-conjunctions heuristics and whether population-based priors improve predictions.

7.1. Tests of Upper Bounds on Parameter Recovery (No Response Errors)

Figure 3 presents key results. To simplify interpretation we plot random queries in Appendix D, rather than Figure 3, because the results are indistinguishable from market-based queries on the scale of Figure 3. Market-based queries do approximately 3% better than random queries for the first 16 queries, approximately 1% better for the first 32 queries, and approximately 0.5% better for all 256 queries. (Queries 129–256 are not shown in Figure 3; the random-query and the market-based query curves asymptote to 100%.) Orthogonal-design questions are only defined for $K = 32$.

Questions selected adaptively by the active-learning algorithm find respondents' decision heuristics much more rapidly than existing question-selection methods. The adaptive questions come very close to an optimal reduction in posterior entropy. 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 (four bits). Each subsequent query is a binary outcome that can reveal at most one bit. A perfect active-learning algorithm would

Figure 3 **Synthetic Data Experiments (Base Comparison, No Error): Percent Uncertainty Explained (U^2) for Alternative Question-Selection Methods**



require 12 additional queries to identify a decision rule (4 bits + 12 bits identifies the 16 elements of $\vec{a}_i$). On average, in the absence of response error, the adaptive questions identify the respondent's decision heuristic in approximately 13 questions. The variational approximation and the one-step-ahead question-selection algorithm appear to achieve close-to-optimal information (12 bits in 13 questions).

We compare the relative improvement as a result of question-selection methods by holding information constant and examining how many questions it takes to achieve that level of parameter recovery. Because an orthogonal design is fixed at 32 questions, we use it as a benchmark. As illustrated in the first line of data in Table 1, an orthogonal design requires 32 queries to achieve a U^2 of approximately 76%. Market-based questions require 38 queries; random questions require 40 queries, and adaptive questions only nine queries. To parse the configurator from the adaptive questions, Appendix D plots the U^2 obtained with a configurator plus market-based questions. The plot parallels the plot of purely market-based queries requiring 30 queries to achieve a U^2 of approximately 76%. In summary, in an errorless world, the active-learning algorithm chooses adaptive questions that provide substantially more information per question than existing nonadaptive methods. The large improvements in U^2, even for small numbers of questions, suggests that adaptive questions are chosen to provide information efficiently.

7.2. Tests of Parameter Recovery When There Are Response Errors or "Bad" Priors

We now add either response error or "bad" priors and repeat the synthetic data experiments. The plots remain quasi-concave for a variety of levels of response error and/or bad priors.[2] We report representative values in Table 1. (Table 1 is based on false negatives occurring 5% of the time. False positives are set by the corresponding η. Bad priors perturb "good" priors with bias drawn from U[0, 0.1].) Naturally, as we add errors or bad priors, the amount of information obtained per question decreases; for example, 13 adaptive questions achieved a U^2 of 100% without response errors but only 55.5% with response errors. On average, it takes 12.4 adaptive questions to obtain a U^2 of 50% (standard deviation 8.7). The last column of Table 1 reports the information obtained by 32 orthogonal questions. Adaptive questions obtain relatively more information per question than existing methods under all scenarios. Indeed, adaptive questions appear to be more robust to bad priors than existing question-selection methods.

7.3. Tests of the Ability to Recovery Disjunctions-of-Conjunctions Heuristics

We now generate synthetic data for respondents who have two distinct conjunctions rather than just one conjunction. By distinct, we mean no overlap in the conjunctions. We allow both question-selection methods to allocate one-half of their questions to the first conjunction and one-half to the second conjunction. To make the comparison fair, all question-selection methods use data from the two configurators when estimating the parameters of the disjunctions-of-conjunctions heuristics. After 32 questions (plus two configurators), estimates based on adaptive questions achieve a U^2 of 80.0%, whereas random questions achieve a U^2 of only 34.5%. Adaptive questions also beat market-based and orthogonal-design questions handily.

This is an important result. With random questions false positives from the second conjunction pollute the estimation of the parameters of the first conjunction, and vice versa. The active-learning algorithm focuses questions on one or the other conjunction to provide good recovery of the parameters of both conjunctions. We expect the two-conjunction results to extend readily to more than two conjunctions. Although this initial test is promising, future tests might improve the algorithm with endogenous stopping rules that allocate questions optimally among conjunctions.

7.4. Tests of Incorporating Data from Previous Respondents

To demonstrate the value of incorporating data from other respondents, we split the sample of synthetic respondents into two halves. For the first half of the sample, we use bad priors, ask questions adaptively, and estimate the $\vec{a}_i$s. We use the estimated $\vec{a}_i$s and a collaborative filter on two features to customize priors for the remaining respondents. We then ask questions of the remaining respondents using

[2] Although the plots in Figure 2 are concave, there is no guarantee that the plots remain concave for all situations. However, we do expect all plots to be quasi-concave, and they are.

Table 1 **Synthetic Data Experiments Number of Questions Necessary to Match Predictive Ability of 32 Orthogonal Questions**

	Adaptive questions	Random questions	Market-based questions	Orthogonal-design questions	Percent uncertainty[a]
Base comparison	9	40	39	32	76.1
Error in answers	11	38	38	32	53.6
“Bad” priors	6	42	41	32	50.4

Note. Number of questions in addition to the configurator question.

[a] U^2 (percent uncertainty explained) when heuristics estimated from 32 orthogonal questions; U^2 for other question-selection methods is approximately the same subject to integer constraints on the number of questions.

collaborative-filter-based priors. On average, U^2 is 17.8% larger on the second set of respondents (using collaborative-filter-based priors) than on the first set of respondents (not using collaborative-filter-based priors). Thus, even when we use bad priors for early respondents, the posteriors from those respondents are sufficient for the collaborative filter. The collaborative-filter-based priors improve U^2 for the remaining respondents.

7.5. Summary of Synthetic Data Experiments

The synthetic data experiments suggest that

- adaptive question selection via active learning is feasible and can recover the parameters of known heuristic decision rules,
- adaptive question selection provides more information per question than existing methods,
- one-step-ahead active-learning adaptive questions achieve gains in information (reduction in entropy) that are close to the theoretical maximum when there are no response errors,
- adaptive question selection provides more information per question when there are response errors,
- adaptive question selection provides more information per question when there are badly chosen priors,
- it is feasible to extend adaptive-question selection to disjunctions-of-conjunctions heuristic decision rules, and
- incorporating data from other respondents improves parameter recovery.

These synthetic data experiments establish that if respondents use heuristic decision rules, then the active-learning algorithm provides a means to ask questions that provide substantially more information per question.

8. Illustrative Empirical Application with a Large Number of Aspects

In the spring of 2009, a large American automotive manufacturer (AAM) recognized that consideration of their vehicles was well below that of non-U.S. vehicles. Management was interested in exploring various means to increase consideration. As part of that effort, AAM fielded a Web-based survey to 2,336 respondents recruited and balanced demographically 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 AAM’s survey explored various marketing strategies that AAM might use to enhance consideration of their brands. The managerial test of communications strategies is tangential to the scope and focus of this paper, but we illustrate in §10 the types of insight provided by estimating consumers’ noncompensatory heuristics.

Because AAM’s managerial decisions depended on the accuracy with which they could evaluate their communications strategies, we were given the opportunity to test adaptive-question selection for a subset of the respondents. A subset of 872 respondents was not shown any communications inductions. 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 AAM.) 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).

8.1. Adaptive Question Selection for Calibration Profiles

To test adaptive question selection, one-half of the calibration profiles were chosen adaptively 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 United States. To avoid order effects and to introduce variation in the data, the question-selection methods were randomized. This probabilistic variation means that the number of queries of each type is 14.5, on average, but varies by respondent.

As a benchmark we chose market-based queries rather than random queries. The market-based queries perform slightly better on synthetic data 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 a 53-aspect orthogonal design. (A full factorial would require 357,210 profiles.) 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. Following Sándor and Wedel (2002) and Vriens et al. (2001), we split the market-based profiles (randomly) over respondents.

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 adaptive queries chosen by the active-learning algorithm. They might prevent the algorithm from getting stuck in a local maximum (an analogy to simulated annealing).

8.2. Selecting Priors for the Empirical Application

AAM had co-occurrence data available from prior research, so we set priors as described in §6.1. In addition, using AAM's data and managerial beliefs, we were able to set priors on some pairwise conjunctions such as "Porsche $\wedge$ Kia" and "Porsche $\wedge$ pick-up." Rather than setting these priors directly as correlations among the a_{ij}s, AAM's managers found it more intuitive to generate "pseudo-questions" in which the respondent was assumed to "not consider" a "Porsche $\wedge$ pick-up" with probability q, where q was set by managerial judgment. In other applications researchers might set the priors directly.

8.3. Validation Profiles Used to Evaluate Predictive Ability

After a memory-cleansing task, respondents were shown a second set of 29 profiles, this time chosen by the market-based question-selection method. 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 ϵs) appears to be appropriate.

8.4. Performance Measures

Although hit rate is an intuitive measure, it can mislead intuition 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%. Such a null model, however, provides no information, has a large number of false-negative predictions, and predicts a consideration set size of 0. 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 it would achieve a lower hit rate (68%: $0.68 = (0.8)^2 + (0.2)^2$). Nonetheless, for interested readers, Appendix E provides hit rates.

We expand evaluative criteria by examining 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. However, without knowing specific loss functions to weigh false positives and false negatives differently, we cannot have a single managerial criterion (e.g., Toubia and Hauser 2007). Fortunately, information theory provides a commonly used measure that balances false positives and false negatives: the Kullback–Leibler divergence (KL). KL is a nonsymmetric measure of the difference 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 C provides formulae for the KL measure appropriate to the data in this paper. We calculate divergence from perfect prediction; hence *a smaller KL is better*.

In synthetic data we knew the "true" decision rule and could compare the estimated parameters a_{ij}s to known parameters. U^2 was the appropriate measure. With empirical data we do not know the true decision rule; we only observe the respondents' judgments about consider versus not consider; hence KL is an appropriate measure. However, both attempt to quantify the information explained by the estimated parameters (decision heuristics).

8.5. Key Empirical Results

Table 2 summarizes KL divergence for the two question-selection methods that we tested: adaptive questions and market-based questions. For each question type, we use two estimation methods: (1) the variational Bayes belief-propagation algorithm computes the posterior distribution of the noncompensatory heuristics, and (2) a hierarchical Bayes logit model (HB) computes the posterior distribution for a compensatory model. HB is the most used estimation method for additive utility models (Sawtooth 2004), and it has proven accurate for zero-versus-one consideration decisions (Ding et al. 2011, Hauser et al. 2010).

Table 2 Illustrative Empirical Application KL Divergence for Question-Selection-and-Estimation Combinations (Where Smaller Is Better)

	Noncompensatory heuristics	Compensatory decision model
Question-selection method		
Adaptive questions	0.475[abc]	0.537[c]
Market-based questions	0.512[c]	0.512[cd]
Null models		
Consider all profiles	0.565	
Consider no profiles	0.565	
Randomly consider profiles	0.562	

[a] Significantly better than market-based questions for noncompensatory heuristics ($p < 0.001$).

[b] Significantly better than compensatory decision model ($p < 0.001$).

[c] Significantly better than null models ($p < 0.001$).

[d] Significantly better than adaptive questions for compensatory decision model ($p < 0.001$).

The latter authors provide a full HB specification in Appendix C. Both estimation methods are based only on the calibration data. For comparison Table 2 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.

When the estimation method assumes respondents use heuristic decision rules, rules estimated from adaptive questions predict significantly better than rules estimated from market-based queries. (Hit rates are also significantly better.) Furthermore, for adaptive questions, heuristic rules predict significantly better than HB-estimated additive rules. Although HB-estimated additive models nest lexicographic models (and hence conjunctive models for consideration data), the required ratio of partworths is approximately 10^{15} and not realistic empirically. More likely, HB does less well because its assumed additive model with 53 parameters overfits the data, even with shrinkage to the population mean.

It is perhaps surprising that ~14.5 adaptive questions do so well for 53 aspects. This is an empirical issue, but we speculate that the underlying reasons are (1) consumers use cognitively simple heuristics with relatively few aspects, (2) the adaptive questions search the space of decision rules efficiently to confirm the cognitively simple rules, (3) the configurator focuses this search quickly, and (4) consumer-specific priors keep the search focused.

There is an interesting, but not surprising, interaction effect in Table 2. If the estimation assumes an additive model, noncompensatory-focused adaptive questions do not do as well as market-based questions. Also, consistent with prior research using nonadaptive questions (e.g., Dieckmann et al. 2009, Kohli and Jedidi 2007, Yee et al. 2007), noncompensatory estimation is comparable to compensatory estimation using market-based questions. Perhaps to truly identify heuristics, we need heuristic-focused adaptive questions.

But are consumers compensatory or noncompensatory? The adaptive-question–noncompensatory-estimation combination is significantly better than all other combinations in Table 2. But what if we estimated both noncompensatory and compensatory models using all 29 questions (combining ~14.5 adaptive questions and ~14.5 market-based questions)? The noncompensatory model predicts significantly better than the compensatory model when all 29 questions are used (KL = 0.451 versus KL = 0.560, $p < 0.001$ using a paired t-test). Differences are also significant at $p < 0.001$ using a related-samples Wilcoxon signed-rank test. Because we may not know a priori whether the respondent is noncompensatory or compensatory, collecting data both ways gives us flexibility for post-data-collection reestimation. (In the automotive illustration, prior theory suggested that consumers were likely to use noncompensatory heuristics.)

8.6. Summary of Empirical Illustration

Adaptive questions to identify noncompensatory heuristics are promising. We appear able to select questions to provide significantly more information per query than market-based queries. Furthermore, it appears that questions are chosen efficiently because we can predict well with ~14.5 questions, even in a complex product category with 53 aspects. This is an indication of cognitive simplicity. Finally, consumers appear to be noncompensatory.

9. Initial Tests of Generalizations: Disjunctions of Conjunctions and Population Priors

Although data were collected based on the conjunctive active-learning algorithm, we undertake exploratory empirical tests of two proposed generalizations: disjunctions of conjunctions and prior-respondent-based priors. These exploratory tests complement the theory in §§6.5 and 6.6 and the synthetic data tests in §§7.3 and 7.4.

9.1. Disjunctions of Conjunctions

AAM managers sought to focus on consumers' primary conjunctions because, in prior studies sponsored by AAM, 93% of the respondents used only one conjunction (Hauser et al. 2010). However, we might gain additional predictive power by searching for second and subsequent conjunctions using the methods of §6.5. Ideally, this requires new data, but we get an indicator by (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 because there were more conjunctions. It came at the expense of a slight increase in false positives. Overall, using all 29 questions, KL increased slightly (0.459 versus 0.452, $p < 0.001$), suggesting that the reestimation on incorrectly classified profiles overfit the data. Because the disjunctions of conjunctions (DOC) generalization works for synthetic data, a true test awaits new empirical data.

9.2. Previous-Respondent-Based Priors

The priors used to initialize consumer-specific beliefs were based on judgments by AAM managers and analysts; however, we might also use the methods proposed in §6.6 to improve priors based on data from other respondents. As a test, we used the basic algorithm to estimate the p_{ijK}s, used the collaborative filter to reset the priors for each respondent, reestimated the model (p'_{ijK}s), and compared predicted consideration to observed consideration. Previous-respondent-based priors improved predictions but not significantly (0.448 versus 0.452, $p = 0.082$), suggesting that AAM provided good priors for this application.

10. Managerial Use

The validation reported in this paper was part of a much larger effort by AAM to identify communications strategies that would encourage consumers to consider AAM vehicles. At the time of the study, two of the three American manufacturers had entered bankruptcy. AAM's top management believed that overcoming consumers' unwillingness to consider AAM vehicles was critical if AAM was to become profitable. Table 2, combined with ongoing studies by AAM, was deemed sufficient evidence for managers to rely on the algorithm to identify consumers' heuristic decision rules. AAM is convinced of the relevancy of consumer heuristics and is actively investigating how to use noncompensatory data routinely to inform management decisions. We summarize here AAM's initial use of information on consumer heuristics.

The remaining 1,464 respondents each answered 29 adaptive plus market-based questions, were shown an experimental induction, and then answered a second set of 29 adaptive plus market-based questions. Each induction was a communications strategy targeted to influence consumers to (1) consider AAM vehicles or (2) consider vehicles with aspects on which AAM excelled. Details are proprietary and beyond the scope of this paper. However, in general, the most effective communications strategies were those that surprised consumers with AAM's success in a non U.S. reference group. AAM's then-current emphasis on J.D. Power and *Consumer Reports* ratings did not change consumers' decision heuristics.

AAM used the data on decision heuristics for product development. AAM recognized heterogeneity in heuristics and identified clusters of consumers who share decision heuristics. There were four main clusters: high selectivity on brand and body type, selectivity on brand, selectivity on body type, and (likely) compensatory. There were two to six subclusters

Table 3 **Percentage of Respondents Using Aspect as an Elimination Criterion**

Brand	Elimination (%)	Body type	Elimination (%)	Engine type	Elimination (%)
BMW	68	Sports car	84	Gasoline	3
Buick	97	Hatchback	81	Hybrid	44
Cadillac	86	Compact sedan	62	Engine power	
Chevrolet	34	Standard sedan	58	4 cylinders	9
Chrysler	66	Crossover	62	6 cylinders	11
Dodge	60	Small SUV	61	8 cylinders	69
Ford	23	Full-size SUV	71	EPA rating	
GMC	95	Pickup truck	82	15 mpg	79
Honda	14	Minivan	90	20 mpg	42
Hyundai	89	Quality		25 mpg	16
Jeep	96	Q-rating 5	0	30 mpg	5
Kia	95	Q-rating 4	1	35 mpg	0
Lexus	86	Q-rating 3	23	Price ($)	
Lincoln	98	Crash test		12,000	77
Mazda	90	C-rating 5	0	17,000	54
Nissan	14	C-rating 4	27	22,000	46
Pontiac	97	C-rating 3	27	27,000	48
Saturn	95			32,000	61
Subaru	99			37,000	71
Toyota	15			45,000	87
VW	86				

within each main cluster, for a total of 20 clusters.[3] Each subcluster was linked to demographic and other decision variables to suggest directed communications and product development strategies. Decision rules for targeted consumer segments are proprietary, but the population averages are not. Table 3 indicates which percentage of the population uses elimination rules for each of the measured aspects.

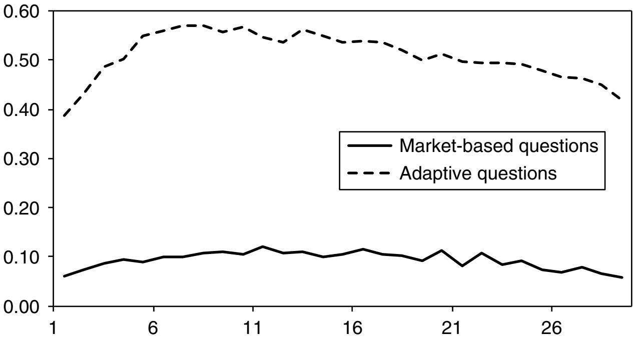
Although some brands were eliminated by most consumers, larger manufacturers have many targeted brands. For example, Buick was eliminated by 97% of the consumers and Lincoln by 98%, but these are not the only GM and Ford brands. For AAM, the net consideration of its brands was within the range of more-aggregate studies. Consumers are mixed on their interest in "green" technology: 44% eliminate hybrids from consideration, but 69% also eliminate large engines. Price elimination illustrates that heuristics are screening criteria, not surrogates for utility: 77% of consumers will not investigate a $12,000 vehicle. This means that consumers' knowledge of the market tells them that, net of search costs, their best strategy is to avoid investing time and effort to evaluate $12,000 vehicles. It does not mean that consumers would not buy a top-of-the-line Lexus if it were offered for $12,000. Table 3 provides aggregate summaries across many consumer segments—AAM's product development and communications strategies were targeted within segment. For example, 84% of consumers overall eliminate sports cars indicating the sports-car segment is a relatively small market. However, the remaining 16% of consumers constitute a market that is sufficiently large for AAM to target vehicles for that market.

11. Summary and Challenges

We found active machine learning to be an effective methodology to select questions adaptively in order to identify consideration heuristics. Both the synthetic data experiments and the proof-of-concept empirical illustration are promising, but many challenges remain.

Question selection might be improved further with experience in choosing "tuning" parameters (ϵ's, T), improved priors, an improved focus on more-complex heuristics, and better variational Bayes belief-propagation approximations. In addition, further experience will provide insight on the information gained as the algorithm learns. For example, Figure 4 plots the average expected reduction in entropy for adaptive questions and for market-based questions. We see that, on average, adaptive questions provide substantially more information per question (5.5 times as much). Prior to the 10th question, the increasingly accurate posterior probabilities enable the algorithm to ask increasingly more accurate questions. Beyond 10 questions the expected reduction in entropy decreases and continues to decrease through the 29th question. It is likely that AAM would have been better able to identify consumers' conjunctive decision rules had they used 58 questions for estimation rather than split the questions between calibration and validation. Research might explore the mix between adaptive and market-based questions.

Figure 4 Average Expected Reduction in Entropy up to the 29th Question



The likelihood principle implies that other models can be tested on AAMs and other adaptive data. The variational Bayes belief-propagation algorithm does not estimate standard errors for the p_{ij}s. Other Bayesian methods might specify more complex distributions. Reestimation or bootstrapping, when feasible, might improve estimation.

Active machine learning might also be extended to other data-collection formats, including formats in which multiple profiles are shown on the same page or formats in which configurators are used in creative ways. The challenge for large N is that we would like to approximate decision rules in less than N queries per respondent.

Acknowledgments

This research was supported by the MIT Sloan School of Management, the Center for Digital Business at MIT (http://ebusiness.mit.edu), and an unnamed American automotive manufacturer. The authors thank Rene Befurt, Sigal Cordeiro, Theodoros Evgeniou, Vivek Farias, Patricia Hawkins, Dmitriy Katz, Phillip Keenan, Andy Norton, Ele Ocholi, James Orlin, Erin MacDonald, Daniel Roesch, Joyce Salisbury, Matt Selove, Olivier Toubia, Paul Tsier, Catherine Tucker, Glen Urban, Kevin Wang, and Juanjuan Zhang for their insights, inspiration, and help on this project. Both reviewers and the area editor provided excellent comments that improved the paper.

[3] AAM used standard clustering methods on the posterior p_{ij}s. By the likelihood principle, it is possible to use latent-structure models to reanalyze the data. Post hoc clustering is likely to lead to more clusters than latent-structure modeling. Comparisons of clustering methods are beyond the scope and tangential to our current focus on methods to select questions efficiently for the estimation of heuristic decision rules.

Appendix A. Example Where Uncertainty Sampling Minimizes Posterior Entropy

We choose a simple example with two aspects to demonstrate the intuition. For this formal example, 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, $\vec{a}_i = \{-1, -1\}, \{-1, +1\}, \{+1, -1\}$, and $\{+1, +1\}$, each of which is a priori equally likely. However, 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 $\vec{a}_i = \{-1, -1\}$. 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	$-2\log_2\frac{1}{2} = 2$

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 B. Pseudo-Code for Belief-Propagation Algorithm

Maintain the notation of the text; let $\vec{p}_{iK}$ be the vector of the p_{ijK}s, and let $\vec{p}_{iK,-j}$ be the vector of all but the jth element. Define two index sets, $S_j^+ = \{k \mid x_{ijk} = 1, y_{ik} = 1\}$ and $S_j^- = \{k \mid x_{ijk} = 1, y_{ik} = 0\}$. 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 Kth query. In application, the ϵs 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 $\Pr(y_{ik} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = \pm 1)$.
While $\max_j (p_{ijK}^h - p_{ijK}^{h-1}) > 0.001$.        [Continue looping until $p_{ijK}^h$ converges.]
    For $j = 1$ to $N$        [Loop over all aspects.]
        For $k \in S_j^+$        [Use variational distribution to approximate data likelihood.]
            $\Pr(y_{ik} = 1 \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = 1) = (1-\epsilon_2)\prod_{x_{igk}=1,\, g\neq j} p_{igK}^{h-1} + \epsilon_1\big(1 - \prod_{x_{igk}=1,\, g\neq j} p_{igK}^{h-1}\big)$,
            $\Pr(y_{ik} = 1 \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = -1) = \epsilon_1$
        end loop $k \in S_j^+$
        For $k \in S_j^-$        [Use variational distribution to approximate data likelihood.]
            $\Pr(y_{ik} = 0 \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = 1) = (1-\epsilon_1)\big(1 - \prod_{x_{igk}=1,g\neq j} p_{igK}^{h-1}\big) + \epsilon_2 \prod_{x_{igk}=1,g\neq j} p_{igK}^{h-1}$,
            $\Pr(y_{ik} = 0 \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = -1) = (1-\epsilon_1)$
        end loop $k \in S_j^-$
        $\Pr(\vec{y}_{iK} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij}=1) = \prod_{k=1}^{K} \Pr(y_{ik} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij}=1)$,
        $\Pr(\vec{y}_{iK} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij}=-1) = \prod_{k=1}^{K} \Pr(y_{ik} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij}=-1)$
        [Compute data likelihoods across all $K$ questions as a product of marginal distributions for each $k$.]
        $\Pr(a_{ij} = 1 \mid X_{iK}, \vec{y}_{iK}, \vec{p}_{iK,-j}^{h-1}) \propto \Pr(\vec{y}_{iK} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = 1)\Pr(a_{ij} = 1 \mid prior)$,
        $\Pr(a_{ij} = -1 \mid X_{iK}, \vec{y}_{iK}, \vec{p}_{iK,-j}^{h-1}) \propto \Pr(\vec{y}_{iK} \mid X_{iK}, \vec{p}_{iK,-j}^{h-1}, a_{ij} = -1)(1 - \Pr(a_{ij} = 1 \mid prior))$,
        $p_{ijK}^h = \Pr(a_{ij} = 1 \mid X_{iK}, \vec{y}_{iK}, \vec{p}_{iK,-j}^{h-1})$ normalized.
                                        [Use Bayes theorem, then normalize.]
    end loop $j$        [Test for convergence and continue if necessary.]
```

Appendix C. Kullback–Leibler Divergence for Empirical 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 (to consider or not). For respondent i we predict that profile k is considered with probability, $r_{ik} = \Pr(y_{ik} = 1 \mid \vec{x}_{ik}, model)$. Then the divergence from the true model (the y_{ik}s) to the model being tested (the r_{ik}s) is given by Equation (C1). With log-based-2, KL has the units of bits:

$$KL = \sum_{k \in \text{validation}} \left[y_{ik} \log_2\left(\frac{y_{ik}}{r_{ik}}\right) + (1 - y_{ik}) \log_2\left(\frac{1 - y_{ik}}{1 - r_{ik}}\right)\right]. \quad \text{(C1)}$$

When the r_{ik}s are themselves discrete, we must 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, let $\hat{C}_v =$ the number of considered validation profiles, let $F_p =$ the false-positive predictions, and let $F_n =$ the false-negative predictions. Then KL 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):

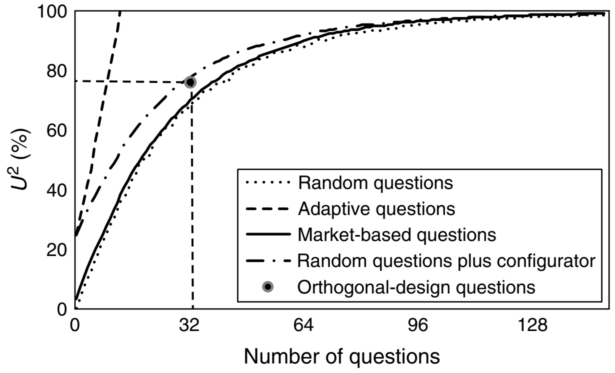
$$KL = \sum_{S_{c,c}} \log_2\left(\frac{\hat{C}_v}{\hat{C}_v - F_p}\right) + \sum_{S_{c,nc}} \log_2\left(\frac{V-\hat{C}_v}{F_n}\right) + \sum_{S_{nc,c}} \log_2\left(\frac{\hat{C}_v}{F_p}\right) + \sum_{S_{c,c}} \log_2\left(\frac{V-\hat{C}_v}{V-\hat{C}_v-F_n}\right).$$

After algebraic simplification, KL can be written as

$$KL = \hat{C}_v \log_2 \hat{C}_v + (V-\hat{C}_v)\log_2(V-\hat{C}_v) - (\hat{C}_v - F_p)\log_2(\hat{C}_v - F_p) - F_n \log_2 F_n - F_p \log_2 F_p - (V-\hat{C}_v - F_n)\log_2(V-\hat{C}_v-F_n). \tag{C2}$$

KL 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 KL would be twice as large. For comparability across respondents with different validation set sizes, we divide by V to scale KL.

Appendix D. Percent Uncertainty Explained (U^2) for Other Question-Selection Methods



Appendix E. Hit Rates for Question-Selection-and-Estimation Combinations (Where Larger Is Better)

	Noncompensatory heuristics	Compensatory decision model
Question-selection method		
Adaptive questions	0.848[abcde]	0.594[d]
Market-based questions	0.827[bcde]	0.806[cdf]
Null models		
Consider all profiles	0.180	
Consider no profiles	0.820	
Randomly consider profiles	0.732	

[a] Significantly better than market-based questions for noncompensatory heuristics ($p < 0.001$).

[b] Significantly better than the compensatory decision model ($p < 0.001$).

[c] Significantly better than the random null model ($p < 0.001$).

[d] Significantly better than the consider-all-profiles null model ($p < 0.001$).

[e] Significantly better than the consider-no-profiles null model ($p < 0.001$).

[f] Significantly better than adaptive questions for the compensatory decision model ($p < 0.001$).

References

Arora, N., J. Huber. 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *J. Consumer Res.* **28**(2) 273–283.

Attias, H. 1999. Inferring parameters and structure of latent variable models by variational Bayes. K. B. Laskey, H. Prade, eds. *Proc. 15th Conf. Uncertainty Artificial Intelligence* (*UAI-99*), Morgan Kaufmann, San Francisco, 21–30.

Breese, J. S., D. Heckerman, C. Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proc. 14th Conf. Uncertainty Artificial Intelligence* (*UAI-98*), Morgan Kaufmann, San Francisco, 43–52.

Chaloner, K., I. Verdinelli. 1995. Bayesian experimental design: A review. *Statist. Sci.* **10**(3) 273–304.

Dahan, E., J. R. Hauser. 2002. The virtual customer. *J. Product Innovation Management* **19**(5) 332–353.

Dieckmann, A., K. Dippold, H. Dietrich. 2009. Compensatory versus noncompensatory models for predicting consumer preferences. *Judgment Decision Making* **4**(3) 200–213.

Ding, M., J. R. Hauser, S. Dong, D. Dzyabura, Z. Yang, C. Su, S. Gaskin. 2011. Unstructured direct elicitation of decision rules. *J. Marketing Res.* **48**(February) 116–127.

Erdem, T., J. Swait. 2004. Brand credibility, brand consideration, and choice. *J. Consumer Res.* **31**(1) 191–198.

Evgeniou, T., C. Boussios, G. Zacharia. 2005. Generalized robust conjoint estimation. *Marketing Sci.* **24**(3) 415–429.

Frederick, S. 2005. Cognitive reflection and decision making. *J. Econom. Perspect.* **19**(4) 25–42.

Ghahramani, Z., M. J. Beal. 2000. Variational inference for Bayesian mixtures of factor analyzers. *Advances in Neural Information Processing Systems*, Vol 12. MIT Press, Cambridge, MA.

Ghahramani, Z., M. J. Beal. 2001. Propagation algorithms for variational Bayesian learning. *Advances in Neural Information Processing Systems*, Vol. 13. MIT Press, Cambridge, MA.

Gigerenzer, G., D. G. Goldstein. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psych. Rev.* **103**(4) 650–669.

Gigerenzer, G., R. Selten, eds. 2001. *Bounded Rationality: The Adaptive Toolbox*. MIT Press, Cambridge, MA.

Gigerenzer, G., P. M. Todd, The ABC Research Group. 1999. *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford, UK.

Gilbride, T. J., G. M. Allenby. 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Sci.* **23**(3) 391–406.

Gilbride, T. J., G. M. Allenby. 2006. Estimating heterogeneous EBA and economic screening rule choice models. *Marketing Sci.* **25**(5) 494–509.

Green, P. E. 1984. Hybrid models for conjoint analysis: An expository review. *J. Marketing Res.* **21**(2) 155–169.

Green, P. E., V. Srinivasan. 1990. Conjoint analysis in marketing: New developments with implications for research and practice. *J. Marketing* **54**(4) 3–19.

Green, P. E., A. M. Krieger, P. Bansal. 1988. Completely unacceptable levels in conjoint analysis: A cautionary note. *J. Marketing Res.* **25**(August) 293–300.

Hauser, J. R. 1978. Testing the accuracy, usefulness, and significance of probabilistic choice models: An information-theoretic approach. *Oper. Res.* **26**(3) 406–421.

Hauser, J. R., B. Wernerfelt. 1990. An evaluation cost model of consideration sets. *J. Consumer Res.* **16**(4) 393–408.

Hauser, J. R., O. Toubia, T. Evgeniou, R. Befurt, D. Dzyabura. 2010. Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *J. Marketing Res.* **47**(June) 485–496.

Jedidi, K., R. Kohli. 2005. Probabilistic subset-conjunctive models for heterogeneous consumers. *J. Marketing Res.* **42**(4) 483–494.

Kohli, R., K. Jedidi. 2007. Representation and inference of lexicographic preference models and their variants. *Marketing Sci.* **26**(3) 380–399.

Kullback, S., R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.* **22**(1) 79–86.

Lewis, D. D., W. A Gale. 1994. Training text classifiers by uncertainty sampling. *Proc. 17th Annual Internat. ACM SIGIR Conf. Res. Dev. Inform. Retrieval, Dublin, Ireland*, 3–12.

Liberali, G., G. L. Urban, J. R. Hauser. 2011. Field experiments: Providing unbiased competitive information to encourage trust, consideration, and sales. Working paper, MIT Sloan School of Management, Cambridge, MA.

Lindley, D. V. 1956. On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**(4) 986–1005.

Little, J. D. C. 2004a. Managers and models: The concept of a decision calculus. *Management Sci.* **50**(12, Supplement) 1841–1853. [Reprinted from 1970.]

Little, J. D. C. 2004b. Comments on "Models and managers: The concept of a decision calculus": Managerial models for practice. *Management Sci.* **50**(12, Supplement) 1854–1860.

Liu, Q., N. Arora. 2011. Efficient choice designs for a consider-then-choose model. *Marketing Sci.* **30**(2) 321–338.

Martignon, L., U. Hoffrage. 2002. Fast, frugal, and fit: Simple heuristics for paired comparisons. *Theory Decision* **52**(1) 29–71.

Payne, J. W., J. R. Bettman, E. J. Johnson. 1988. Adaptive strategy selection in decision making. *J. Experiment. Psych.: Learn., Memory, Cognition* **14**(3) 534–552.

Payne, J. W., J. R. Bettman, E. J. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge, UK.

Rényi, A. 1961. On measures of information and entropy. *Proc. 4th Berkeley Sympos. Math., Statistics, Probability*, University of California, Berkeley, Berkeley, 547–561.

Sándor, Z., M. Wedel. 2002. Profile construction in experimental choice designs for mixed logit models. *Marketing Sci.* **21**(4) 455–475.

Sawtooth Software. 1996. ACA system: Adaptive conjoint analysis. *ACA Manual*. Sawtooth Software, Sequim, WA.

Sawtooth Software. 2004. CBC hierarchical Bayes analysis. Technical paper, Sawtooth Software, Sequim, WA.

Sawtooth Software. 2008. ACBC. Technical paper, Sawtooth Software, Sequim, WA.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Tech. J.* **27**(July and October) 379–423, 623–656.

Toubia, O., J. R. Hauser. 2007. On managerially efficient experimental designs. *Marketing Sci.* **26**(6) 851–858.

Toubia, O., J. R. Hauser, R. Garcia. 2007. Probabilistic polyhedral methods for adaptive choice-based conjoint analysis: Theory and application. *Marketing Sci.* **26**(5) 596–610.

Toubia, O., J. R. Hauser, D. I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *J. Marketing Res.* **41**(February) 116–131.

Tversky, A. 1972. Elimination by aspects: A theory of choice. *Psych. Rev.* **79**(4) 281–299.

Urban, G. L., J. R. Hauser. 2004. "Listening in" to find and explore new combinations of customer needs. *J. Marketing* **68**(2) 72–87.

van Nierop, E., B. Bronnenberg, R. Paap, M. Wedel, P. H. Franses. 2010. Retrieving unobserved consideration sets from household panel data. *J. Marketing Res.* **47**(February) 63–74.

Vriens, M., M. Wedel, Z. Sándor. 2001. Split-questionnaire designs: A new tool in survey design. *Marketing Res.* **13**(2) 14–19.

Wilkie, W. L., E. A. Pessemier. 1973. Issues in marketing's use of multi-attribute attitude models. *J. Marketing Res.* **10**(4) 428–441.

Yedidia, J. S., W. T. Freeman, Y. Weiss. 2003. Understanding belief propagation and its generalizations. G. Lakemeyer, B. Nebel, eds. *Exploring Artificial Intelligence in the New Millennium*. Morgan Kaufmann, San Francisco, 239–269.

Yee, M., E. Dahan, J. R. Hauser, J. Orlin. 2007. Greedoid-based noncompensatory inference. *Marketing Sci.* **26**(4) 532–549.