The authors test methods, based on cognitively simple decision rules, that predict which products consumers select for their consideration sets. Drawing on qualitative research, the authors propose disjunctions-of-conjunctions (DOC) decision rules that generalize well-studied decision models, such as disjunctive, conjunctive, lexicographic, and subset conjunctive rules. They propose two machine-learning methods to estimate cognitively simple DOC rules. They observe consumers' consideration sets for global positioning systems for both calibration and validation data. They compare the proposed methods with both machine-learning and hierarchical Bayes methods, each based on five extant compensatory and noncompensatory rules. For the validation data, the cognitively simple DOC-based methods predict better than the ten benchmark methods on an information theoretic measure and on hit rates. The results are robust with respect to format by which consideration is measured, sample, and presentation of profiles. The article closes with an illustration of how DOC-based rules can affect managerial decisions.

Keywords: consideration sets, noncompensatory decisions, consumer heuristics, machine learning, conjoint analysis, cognitive simplicity, lexicography, decision theory

Disjunctions of Conjunctions, Cognitive Simplicity, and Consideration Sets

Consideration decisions are managerially important. For example, General Motors has invested heavily in product design and quality; in 2007, Buick tied Lexus for the top spot in J.D. Power and Associates’ vehicle dependability ranking, and in 2008, Buick was the top U.S. brand in Consumer Reports. However, roughly half of U.S. consumers (64% in California) will not even consider a Buick. Because the typical consumer considers fewer than 10 vehicles when shopping for a new vehicle, top managers at General Motors are interested in understanding how consumers decide which 10 of the 350-plus make-model combinations to consider further. To direct strategies, they would like a model that can forecast changes in consideration as a function of changes in product lines or changes in the features that are emphasized in marketing activities.

Two-stage, consider-then-choose decision rules are particularly relevant in the automobile market, but modeling and forecasting such decision rules is of general interest. When consumers face a large number of alternative products, as is increasingly common in today’s retail and Web-based shopping environments, they typically screen the full set of products down to a smaller, more manageable consid-
eration set, which they then evaluate further (e.g., Bronnenberg and Vanhonacker 1996; DeSarbo et al. 1996; Hauser and Wernerfelt 1990; Jedidi, Kohli, and DeSarbo 1996; Mehta, Rajiv, and Srinivasan 2003; Montgomery and Svenson 1976; Payne 1976; Roberts and Lattin 1991; Shocker et al. 1991; Wu and Rangaswamy 2003). Consideration sets for packaged goods typically comprise 3–4 products rather than the 30–40 products on the market (Hauser and Wernerfelt 1990; Urban and Hauser 2004). Forecasting consideration sets can explain approximately 80% of the explainable uncertainty in consumer decision making (assuming equally likely choice within the consideration set; Hauser 1978).

In complex product categories, research suggests that at least some consumers use noncompensatory decision processes when evaluating many products and/or products with many features (e.g., Payne, Bettman, and Johnson 1988, 1993).

In this article, we explore machine-learning algorithms based on noncompensatory decision rules that model consumers’ decisions in the consideration stage of a consider-then-choose process. We measure consideration directly for a moderately complex product, handheld global positioning systems (GPSs), and assuming a general form of noncompensatory decision rules, we attempt to model the noncompensatory patterns that best predict consumers’ consideration decisions. The general form, disjunctions of conjunctions (DOC), is motivated by qualitative data and systems (GPSs), and assuming a general form of noncompensatory patterns that best predict consumers’ consideration sets. In some formats, respondents do not need to evaluate every profile. In other formats, they do not need to evaluate every profile. In some formats, respondents classify each profile as considered or not considered; in other formats, they do not need to evaluate every profile.

We explore situations in which features are described by finitely many levels. Let $j$ index the profiles, $l$ index the levels, $f$ index the features (sometimes called “attributes” in the literature), and $h$ index the respondents. Let $J$, $L$, $F$, and $H$ be the corresponding numbers of profiles, levels, features, and respondents. For ease of exposition only, we do not write $J$, $L$, and $F$ as dependent (e.g., $L_j$). The models and estimation can (and do) handle such dependency, but the notation is cumbersome. Let $x_{jl} = 1$ if profile $j$ has feature $f$ at level $l$; otherwise, $x_{jl} = 0$. Let $x_j$ be the binary vector (of length $L_F$) describing profile $j$. Let $y_{jh} = 1$ if we observe that respondent $h$ considers profile $j$; otherwise, $y_{jh} = 0$. Let $\bar{y}_h$ be the binary vector describing respondent $h$’s consideration decisions.

**Noncompensatory Decision Rules**

Commonly studied noncompensatory rules include disjunctive, conjunctive, lexicographic, elimination-by-aspects, and subset conjunctive rules (e.g., Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005; Montgomery and Svenson 1976; Ordóñez, Benson, and Beach 1999; Payne, Bettman, and Johnson 1988; Yee et al. 2007). Subset conjunctive rules generalize disjunctive and conjunctive rules (Jedidi and Kohli 2005). For consideration decisions, they also generalize lexicographic rules and deterministic elimination-by-aspects because any implied ranking of products by lexicographic feature-level orders is indeterminate if only the consideration decision is observed (Hogarth and Karelaia 2005; Johnson, Meyer, and Ghose 1989; Montgomery and Svenson 1976; Payne, Bettman, and Johnson 1988; Tversky 1972).

**Disjunctive rules.** In a disjunctive rule, a profile is considered if at least one of the features is at an “acceptable” (or satisfactory) level. Let $a_{hfl} = 1$ if level $l$ of feature $f$ is acceptable to respondent $h$; otherwise, $a_{hfl} = 0$. Let $\bar{a}_h$ be the binary vector of acceptabilities for respondent $h$. A disjunctive rule states that respondent $h$ considers profile $j$ if $\bar{a}_h x_j \geq 1$.

**Conjunctive rules.** In a conjunctive rule, a profile is considered if all the features are at an acceptable level. (Conjunctive rules usually assume a larger set of acceptable levels than disjunctive rules, but this is not required.) Because the use in each rule is clear in context, we use the same notation: In a conjunctive rule, respondent $h$ considers profile $j$ if $\bar{a}_h x_j = F$.

**Subset conjunctive rules.** In a subset conjunctive rule, a profile is considered if at least $S$ features are at an acceptable level. Using the same notation, respondent $h$ considers profile $j$ if $\bar{a}_h x_j \geq S$. A conjunctive rule is a special case in which $S = 1$, and because $\bar{a}_h x_j$ can never exceed $F$, a conjunctive rule is a special case in which $S = F$. We denote subset conjunctive rules by Subset($S$). (Subset conjunctive rules are mathematically equivalent to “image-theory” rules in organizational behavior [see Ordóñez, Benson, and Beach 1999].)

**Additive and q-Compensatory Decision Rules**

Perhaps the most pervasively studied decision rules are additive rules. In an additive rule, consumers consider a profile if its “utility” is above some threshold, $T_h$, which accounts for search and processing costs. If $\bar{b}_h$ is the vector of partworths for respondent $h$, then $h$ considers profile $j$ if $\bar{b}_h x_j \geq T_h$. For estimation, we model errors in the decisions. Many researchers have demonstrated that an additive partworth rule can mimic lexicographic, subset conjunctive, and conjunctive rules (e.g., Jedidi and Kohli 2005; Kohli and Jedidi 2007; Olshevsky and Acito 1980; Yee et al. 2007). To explore whether a model might predict better if it is con-
strained to be compensatory, we follow Bröder (2000) and
Yee and colleagues (2007), who specify a q-compensatory
model by constraining the additive model so that no feature’s importance is more than q times as large as another
target’s importance. (Hogarth and Karelaia [2005] and
Martignon and Hoffrage [2002] use related constraints. A feature’s importance is the difference between the maxi-
mum and the minimum partworths for that feature.)

DOC

To study consideration-set decisions, we began with a qua-
litative study that used in-depth interviewing of 39 auto-
mobile consumers who were asked to describe their consid-
eration decisions for 100 real automobiles balanced to mar-
et data. All interviews were videotaped, and the videos were
evaluated by independent judges who were blind to any hypotheses about consumers’ decision rules (Hughes
and Garrett 1990; Perreault and Leigh 1989). Most respondents made consideration decisions rapidly (89% averaged
less than 5 seconds per profile), and most used noncompen-
satory decision rules (76%). Typically, consumers used
conjunctive-like criteria defined on specific levels of fea-
tures. However, some consumers would consider an auto-
mobile if it satisfied at least one of multiple conjunctive cri-
teria (i.e., a disjunction of two or more conjunctions).

For example, the following respondent considered au-
tomobiles that satisfied either of two criteria. The first crite-
rion is clearly conjunctive (good styling, good interior
room, excellent mileage). The second criterion allows for
cars that are “hot rods,” which usually have poor interior
room and poor mileage.

[I would consider the Toyota Yaris because] the styling
is pretty good, lot of interior room, mileage is supposed
to be out of this world.

I definitely [would] consider [the Infinity M-Sedan],
though I would probably consider the G35 before the
“M”. I like the idea of a kind of a hot rod.

Depth interviewing is, by necessity, based on a small
sample. From the sample, we could not determine whether multiple conjunctions were pervasive or limited to a subset of consumers. However, qualitative interviewing in the handheld GPS category also identified some consumers who used multiple conjunctions. A respondent might be willing to consider a GPS with a black-and-white (B&W) screen if the GPS was small and the screen was high resolution, but this same respondent would require a color screen on a large GPS. Such rules can be written as logical patterns: (B&W screen ∧ small size ∧ high resolution) ∨ (color screen ∧ large size), where ∧ is the logical “and” and ∨ is the logical “or.” Patterns might also include negations (¬); for example, a consumer might accept a B&W screen as long as the GPS is less than the highest price of $399: (B&W screen ∧ ¬$399).

Formal Definition of DOC Rules

To study this phenomenon further, we formalize these qualitative insights with a class of decision rules that general-
izes previously-proposed rules. Following Tversky
(1972), we define an aspect as a binary descriptor, such as “B&W screen.” A profile either has or does not have an
aspect. A pattern is a conjunction of aspects or their nega-
tions, such as (B&W screen ∧ ¬$399). We define the size,
s, of a pattern as the number of aspects in the pattern. For example, (B&W screen ∧ ¬$399) has size s = 2. If p
indexes patterns, we say that a profile j matches pattern p if
profile j contains all aspects (or negations) in pattern p.

We study rules in which a respondent considers a profile
if the profile matches one or more target patterns. Because
each pattern is a conjunction, these logical rules are disjunc-
tions of conjunctions. These DOC rules generalize disjunc-
tive rules (disjunctions of patterns of size 1), conjunctive
rules (patterns of size F), and subset conjunctive rules (pat-
terns of size S).

Let \( w_{hp} = 1 \) if pattern p is one of the patterns describing
respondent h’s decision rule, and let \( m_{jp} = 1 \) if profile
j matches pattern p. Otherwise, \( w_{hp} \) and \( m_{jp} \) are zero. Let \( \bar{w}_h \)
and \( \bar{m}_h \) be the corresponding binary vectors with length
equal to the number of allowable patterns in a DOC rule. A
DOC rule implies that respondent h considers profile j if and
only if \( \bar{m}_j^T \bar{w}_h \geq 1 \).

Cognitive Simplicity

The DOC rules generalize previously proposed noncompen-
satory decision rules, but they might be too general. For
example, any profile can be described by a pattern of size F.
Thus, any consideration set of size n can be fit perfectly
with a disjunction of n conjunctions of size F. Fortunately,
experimental evidence suggests that consumers make con-
sideration decisions with relatively simple rules that enable
them to make good decisions while avoiding excess cogni-
tive effort (e.g., Bettman, Luce, and Payne 1998; Bröder
2000; Gigerenzer and Goldstein 1996; Gigerenzer, Todd,
and the ABC Research Group 1999; Hogarth and Karelaia
2005; Martignon and Hoffrage 2002; Payne, Johnson,
and Bettman 1988, 1993; Shugan 1980; Simon 1955). This perspec-
tive of simple, efficient, search-and-evaluation rules is
consistent with economic theories of consideration-set for-
mation that posit that consumers balance search costs with the
option value of utility maximization (Hauser and Wern-
erfelt 1990; Roberts and Lattin 1991). To capture this “cog-
nitive simplicity,” we define DOC(S) rules as the set of
DOC rules with maximum pattern length S. In addition, we
either limit the number of patterns, P, or penalize DOC rules
that have large P.

MACHINE-LEARNING APPROACHES TO IDENTIFY
DOC PATTERNS

For a set of respondents and profiles, the basic data we
observe are whether a respondent considers a profile \( (y_{ij}) \).
We attempt to identify the patterns that predict best how
respondent h evaluates profiles. Using a calibration sample,
we seek patterns such that \( m_j^T w_h \geq 1 \) whenever we observe
that profile j is considered and such that \( m_j^T w_h = 0 \) when-

---

1In the Web Appendix (http://www.marketingpower.com/jmrjune10), we
formally demonstrate that (1) disjunctive rules, subset conjunctive rules of
pattern length 1, and DOC rules of maximum pattern length 1 are equiva-
 lent; (2) conjunctive rules and subset conjunctive rules of pattern length F
are equivalent and a subset of the DOC rule; and (3) subset conjunctive
rules of pattern length S can be written as DOC rules, but there are DOC
rules of maximum pattern length S that cannot be written as subset con-
jective rules of pattern length S.
ever we observe that profile $j$ is not considered. (Recall that $\tilde{m}_h$ and $\hat{w}_h$ are binary.)

The number of allowable DOC(S) patterns grows rapidly with $S$. For example, with the 16 binary features in the empirical test, there would be 32 patterns for $S = 1$, 512 for $S = 2$, 4,992 for $S = 3$, and 34,112 for $S = 4$. There would be approximately 20 million patterns of length $S = 10$. With only 32 binary observations (consider versus not consider), there is serious concern about overfitting because the vector, $\hat{w}_h$, which we attempt to estimate, has a length equal to this large number of allowable patterns.

Machine learning is particularly suited to this pattern-matching task. Qualitative interviews suggest that it is not unreasonable for patterns to be up to length $S = 4$, which requires that we search more than 34,000 patterns to find those that best fit the data. Although we might place priors on each pattern and use Bayesian methods, we have not yet been able to develop a Bayesian representation in which the posterior is robust with respect to exogenously set priors for the large number of parameters. We leave exploration of Bayesian DOC models to further research.

Rather than producing posterior probabilities of pattern inclusion, we seek binary indicators of whether a pattern is in the best-fit solution. If the data are too noisy or the solution space is too large (even if we control for cognitive simplicity), predictions could overfit the data and predict poorly. To be sensitive to this concern, we compare models using predictive tests in which respondents face an entirely new set of profiles and report consideration for those profiles.

\section*{Cognitive Simplicity and Complexity Control}

Although we used cognitive simplicity to motivate small $S$ and $P$, such constraints or penalties have an alternative interpretation within machine learning—namely, complexity control (e.g., Cucker and Smale 2002; Evgeniou, Bous sios, and Zacharia 2005; Hastie, Tibshirani, and Friedman 2003; Langley 1996; Vapnik 1998). Limiting the complexity of a model often reduces in-sample overfitting and enhances out-of-sample prediction. Both the behavioral explanation and the complexity-control motivation are consistent with the DOC(S) models—we cannot rule out either with the data in this study.

\section*{Sample Shrinkage}

To further distinguish among potential patterns, we use data from the entire sample to help select patterns for respondent $h$. In an analogy to shrinkage, which enhances accuracy in HB models (e.g., Rossi and Allenby 2003), we favor patterns that fit the largest subset of respondents. Although shrinkage alone is sufficient motivation for use in our models, shrinkage is consistent with behavioral theories that suggest that simple rules have evolved because they work well in the general environment in which a sample of consumers often make decisions (e.g., Chase, Hertwig, and Gigerenzer 1998). These researchers hypothesize that consumers continue to use similar (simple) rules when faced with new decisions.

We now briefly summarize two machine-learning methods. Detailed equations are available in the Web Appendix (http://www.marketingpower.com/jmrjune10).

\section*{Mathematical Programming (DOCMC)}

Because we seek the binary vector, $\hat{w}_h$, that best matches patterns in the calibration data, we formulate an integer program such that $w_{ip}$ must be either 0 or 1 for all $p$. For respondent $h$, we define false positives, $FP_h(\hat{w}_h)$, as the number of profiles predicted to be considered but observed as not considered, and we define false negatives, $FN_h(\hat{w}_h)$, as the number of profiles predicted to be not considered but observed to be considered. In its most basic form, the integer program (DOCMC) would choose the $\hat{w}_h$ that minimizes the sum of false positives and false negatives for respondent $h$.

We enforce cognitive simplicity (complexity control) by limiting the search to patterns of length $S$ or less and by penalizing pattern length, $P$. We include shrinkage with terms proportional to the sum of false positives and false negatives in the sample (sum over all respondents). Formally, the objective function is as follows:

\[ \min_{\{\hat{w}_h\}} \left\{ FP_h(\hat{w}_h) + FN_h(\hat{w}_h) + \gamma M \sum_{i=1}^{n} \left[ FP_i(\hat{w}_h) + FN_i(\hat{w}_h) \right] + \gamma P \right\} \]

Note that DOCM is equivalent to a set-covering problem and, thus, is an NP-hard problem (Cormen et al. 2001). Fortunately, efficient greedy approximation algorithms have been developed and tested for this class of problems (Feige 1998; Lund and Yannakakis 1994). Alternatively, DOCM can be solved approximately with a linear-programming relaxation in which we first allow $\hat{w}_h$ to be continuous on $[0, 1]$ and then round up any positive $w_{ih}$ that is above a threshold (see Hastie, Tibshirani, and Friedman 2003, and the references therein). In our estimations, we use both the greedy and the relaxation methods, choosing the solution that provides the best value of the objective function (using calibration data only, no data from the validation profiles).

The DOCM requires three exogenous parameters: $\gamma_M$, which tells us how much to penalize a lack of sample-level fit; $\gamma_P$, which tells us how much to penalize the number of patterns; and $S$, which sets the maximum pattern length. One method to select these parameters is leave-one-out cross-validation (e.g., Coill, Winer, and Rados 1987; Efroin and Tibshirani 1997; Evgeniou, Pontil, and Toubia 2007; Hastie, Tibshirani, and Friedman 2003; Kearns and Ron 1999; Kohavi 1995; Shao 1993; Toubia, Evgeniou, and Hauser 2007; Zhang 2003). Specifically, for potential values of the exogenous “tuning” parameters, we leave out one profile from the calibration data, estimate $\hat{w}_h$; predict consideration for the left-out profile; and choose “tuning” parameters to minimize prediction errors on the holdout profiles. (No data from any holdout or validation observations are used in leave-one-out cross-validation.)

In the data, neither the leave-one-out cross-validation nor out-of-sample predictions are particularly sensitive to the choice of “tuning” parameters within ranges that roughly match a priori beliefs. Such robustness is consistent with Evgeniou, Pontil, and Toubia (2007). Specifically, we can choose any $\gamma_M$ that is an arbitrarily small number such that sample-level consideration is used only to break ties among
patterns. For $\gamma_c$, cross-validation (and predictive tests) varies little in the range $\gamma_c \in [1, 4.5]$. Similarly, we can select a cognitively simple $S$ to be within ranges that we observe in qualitative interviews ($S \sim 2, 3, 4$). We report $S = 4$ for ease of exposition.

Logical Analysis of Data (LAD-DOC)

Logical analysis of data (LAD), which attempts to distinguish “positive” events from “negative” events, is another approach to generate patterns (Boros et al. 1997; Boros et al. 2000). We control cognitive simplicity by limiting the search to at most $P$ patterns of at most size $S$. We define positive patterns as patterns that match at least one considered profile but no nonconsidered profile. Following the “bottom-up” approach that Boros and colleagues (2000) describe, we begin by generating minimal patterns of length one that match some considered profiles. If such patterns are not contained in any nonconsidered profile, they are positive patterns. Otherwise, we add aspects to the patterns one by one until we generate positive patterns, or until we reach maximum length ($S$). Next, we use a greedy algorithm to identify up to $P$ positive patterns that best fit the data, breaking ties first by giving preference to shorter patterns and then patterns that are positive most often in the sample. The union of these positive patterns is a DOC rule. Thus, LAD-DOC provides a contrast to DOCMP. It is simpler to formulate and takes less time to run, but it shares the characteristics of selecting patterns that best fit the data subject to cognitive simplicity ($S, P$) and shrinkage (break ties to fit sample-level consideration). A potential weakness is that the implementation of LAD focuses primarily on avoiding false positives (in the calibration data) rather than a combination of false positives and false negatives. For comparability to DOCMP, we set $S = 4$ and $P = 2$, but out-of-sample predictions are comparable for $P \sim 2, 3, 4$ and $S \sim 4$ or 5.

BENCHMARKS

As benchmarks, we choose five decision rules. We estimate these rules with both machine-learning and HB methods. The decision rules are as follows:

1. Additive partworth rules,
2. Additive q-compensatory rules,
3. Disjunctive rules,
4. Conjunctive rules, and
5. Subset conjunctive rules.

The machine-learning estimations use objective functions comparable to Equation 1. For the additive and q-compensatory rules, we penalize the sum of the partworths rather than the number of patterns. Detailed formulations are available in the Web Appendix (http://www.marketingpower.com/jmrjune10).

The HB methods mimic extant methods to the greatest extent possible. For the additive and q-compensatory rules, we use standard HB choice-based conjoint formulations. For the basic format, we follow the formulation of Allenby, Arora, and Ginter (1995). For conjunctive rules, we modify an algorithm that Gilbride and Allenby (2004) developed. The modifications reflect differences in data and generalization ($S = 1$ or $F$ in Gilbride and Allenby [2004]). As data, we observe consideration directly, while it is a latent construct in Gilbride and Allenby’s formulation. To address unordered multilevel features, we do not impose constraints that levels within a feature are ordered. Detailed HB formulations are available in the Web Appendix (http://www.marketingpower.com/jmrjune10).

For the subset conjunctive rules, we select $S = 4$ to be consistent with the DOC rules. (Predictive tests for other values of $S$ are available on request.) In addition to detailed formulations, the Web Appendix (http://www.marketingpower.com/jmrjune10) also contains simulations that compare some of the benchmarks with DOC-based methods on synthetic data.

EMPIRICAL APPLICATION: GPS

We chose to study GPSs because the number of features and brands available is sufficiently large that we might expect some noncompensatory decision rules. Figure 1 illustrates 16 features that consumers use to evaluate handheld GPSs. We chose these features as the most important on the basis of two pretests of 58 and 56 consumers, respectively. Of the features, 10 are represented by text and icons, and the remaining 6 are represented by text and visual cues.

Using the 16 features, we generated an orthogonal design of 32 GPS profiles.4 We then developed four alternative formats to measure consideration. We developed these respondent task formats on the basis of qualitative pretests to approximate the shopping environment for GPSs. Each respondent task format was implemented in a Web-based survey and was pretested extensively with more than 55 potential respondents from the target market. At the end of the pretests, respondents found the tasks easy to understand and reported that the task formats were reasonable representations of the handheld GPS market.

We invited two sets of respondents to complete the Web-based tasks: a representative sample of German consumers who were familiar with handheld GPSs and a U.S.-based student sample. We first describe the results from the primary format using the German sample of representative consumers. We then discuss the other formats, the student sample, and a text-only version.

Figure 2 provides screenshots in English and German for the basic format. A “bull pen” is on the far left. As respondents move their cursor over a generic image in the bull pen, a GPS appears in the middle panel. If respondents click on the generic image, they can evaluate the GPS in the middle panel and decide whether to consider it. If they decide to consider the GPS, its image appears in the right panel. Respondents can toggle between current consideration sets and their current nonconsidered sets. There are many ways

---

4To make the task realistic and to avoid dominated profiles (Johnson, Meyer, and Ghose 1989), we manipulated price as a two-level price increment. Profile prices were based on this increment plus additive feature-based costs. We return to the issue of orthogonal designs at the end of this section.
they can change their mind—for example, putting a GPS back or moving it from the consideration set to the nonconsidered set, or vice versa. In this format, respondents continue until all GPSs are evaluated.

**Figure 2**
CONSIDERATION TASK IN ONE OF THE FORMATS (ENGLISH AND GERMAN)

Before respondents made consideration decisions, they reviewed screens that described GPSs in general and each of the GPS features. They also viewed instruction screens for the consideration task and instructions that encouraged incentive compatibility. Following the consideration task, respondents ranked profiles within the consideration set (these data are not used in this article) and then completed tasks designed to cleanse their memory. These tasks included short brainteaser questions that directed respondents’ attention away from GPSs. Following the memory-cleansing tasks, respondents completed the consideration task a second time but for a different orthogonal set of GPSs. These second consideration decisions are validation data and are not used in the estimation of any rules.

Respondents were drawn from a Web-based panel of consumers maintained by the GfK Group. Initial screening eliminated respondents who had no interest in buying a GPS and no experience using a GPS. Respondents who completed the questionnaire received an incentive of 200 points (or punkte, in accordance with GfK) toward general prizes and were entered into a lottery in which they could win one of the GPSs (plus cash) that they considered. This lottery was designed to be incentive aligned, as in Ding (2007) and Ding, Grewal, and Liechty (2005). (Respondents who completed only the screening questionnaire received 15 punkte.)

In total, 2320 panelists were invited to answer the screening questions. The incidence rate (percentage eligible) was 64%, the response rate was 47%, and the completion rate was 93%. Respondents were reassigned randomly to one of the five task formats (the basic format in Figure 2, three alternative formats, and a text-only format). After eliminating respondents who had null consideration sets or null nonconsidered sets in the estimation task, we retained 580 respondents. The average size of the consideration set (estimation data) for the task format in Figure 2 was 7.8 profiles. There was considerable variation among respondents (SD = 4.8 profiles). The average size of the consideration set in the validation task was smaller (7.2 profiles) but not signifi-
cantly different. Validation consideration-set sizes had an equally large standard deviation (4.8 profiles).

**PREDICTIVE TESTS**

**Criteria to Compare DOCMP, LAD-DOC, and the Benchmarks**

Hit rate is an intuitive measure that is used commonly when comparing predictive ability. However, with average consideration sets at approximately 7.2 of 32 (22.5%), a null model that predicts that no GPSs will be considered will achieve a hit rate of 77.5%. Thus, we follow Srinivasan (1988), Srinivasan and Park (1997), and Payne, Bettman, and Johnson (1993, p. 128) and report the percentage improvement relative to a random-prediction null model. Percentage improvement is a linear transformation of hit rate, but it is easier to interpret.

More critically, the apparent strong performance of “predict nothing considered” suggests that we gain insight with statistics that reward models that actually try to predict consideration. The ability to predict the consideration-set size can reject bad models, but it is not sufficient to evaluate a good model. A null model of random prediction (proportional to calibration consideration-set size) predicts the validation consideration-set size accurately but achieves a low hit rate of 65.3% and provides no useful information (0% relative hit rate improvement).

Instead, we consider a statistic that is sensitive to false positive predictions, false negative predictions, and predicted consideration-set sizes in the validation data. In particular, we use the Kullback-Leibler divergence (K-L), which measures the expected gain in Shannon’s information measure relative to a random model (Chaloner and Verdinelli 1995; Kullback and Leibler 1951; Lindley 1956).5 The K-L percentage is 0% for both the random null model and the “predict nothing considered” null model. It is 100% for perfect prediction. The K-L percentage rewards models that predict the consideration-set size correctly and favors a mix of false positives and false negatives that reflect true consideration sets over those that do not. It discriminates among models even when the hit rates might otherwise be equal. Together, the three statistics—hit rate, relative hit rate improvement, and the K-L percentage—provide a means to assess relative predictive ability (DOC-based models versus the benchmarks).

**Predictive Comparison of DOCMP, LAD-DOC, and the Benchmarks**

Table 1 summarizes the ability of each estimation method to predict consideration for the validation task. Focusing on the comparison of DOC-based models with the benchmarks, we find that DOC-based predictions are best or not significantly different than best on both hit rates and K-L percentage measures and better than all benchmark estimation methods on both measures. Furthermore, LAD-DOC predicts slightly better than DOCP, but the difference is not significant.

Among the benchmarks, the additive-rule models predict well, with the machine-learning version predicting significantly better than the HB version on both hit rate and K-L percentage (t = 2.6, p < .02; t = 3.7, p < .01, respectively). While the DOC-based methods are best or not significantly different than best on all comparisons, the machine-learning additive model is within 1–2 percentage points on hit rate.6 This is consistent with prior results on the robustness of the linear model for empirical data (e.g., Dawes 1979; Dawes and Corrigan 1974) and consistent with the ability of an additive rule to nest some noncompensatory rules.

---


6We find that LAD-DOC is significantly better than the best (machine-learning) additive model on both hit rate and K-L divergence (t = 2.4, p < .02; t = 4.6, p < .01), and DOCP is better, but not quite significantly so, on hit rate and significantly better on K-L divergence (t = 1.9, p = .06; t = 4.1, p < .01). One reason the additive model does less well on the K-L percentage is that it underpredicts the consideration-set size. We examine the predictive ability of the additive model further in the next section.
Empirical Evidence Is Consistent with Cognitive Simplicity

Although DOCMP and LAD-DOC are designed to favor cognitive simplicity, unconstrained estimation could conceivably predict better. We reestimated DOCMP with the $\gamma$ equal to zero and LAD-DOC without the S and P constraints. For both models, the hit rates are significantly better for the penalized/constrained estimation ($p < .01$ versus 75.7% DOCMP without $\gamma$; $p < .01$ versus 80.4% LAD-DOC without constraints, respectively). Cognitive simplicity also improves the K-L percentage, but the improvements are not quite significant ($p < .16$ versus 29.6%; $p = .07$ versus 32.5%, respectively, for unconstrained DOCMP and LAD-DOC). These results are consistent with a hypothesis that predictions improve when cognitive simplicity is enforced, though the marginal significance for K-L percentages suggests that the cognitive simplicity hypothesis is worth further testing in other contexts.

Despite the large number of potential patterns, DOC-based estimation chose relatively simple rules for the data. The LAD-DOC predictions do not improve significantly, and often degrade, as we increase either pattern length ($S$) or the number of patterns ($P$). For DOCMP, 7.1% of the respondents are represented as using two patterns; the remainder are represented with a single pattern. The increased flexibility of the DOC-based estimation methods seems to improve predictive ability relative to alternative noncompensatory rules and their corresponding estimation methods, even though only 7.1% of the respondents are modeled with two patterns.

Sensitivity to Orthogonal Designs

Significant research in marketing exists on efficient experimental designs for choice-based conjoint experiments (Arora and Huber 2001; Huber and Zwerina 1996; Kanninnen 2002; Toubia and Hauser 2007), but we are unaware of any research on efficient experimental designs for consideration decisions or for the estimation of cognitively simple DOC rules. When decisions are made with respect to the full set of 32 profiles, aspects are uncorrelated up to the resolution of the design, and if there were no errors, we should be able to identify DOC patterns accordingly. However, when profiles are removed, aspects may no longer be uncorrelated, and patterns may not be defined uniquely. As a mild test, we reestimated three models—DOCMP, machine-learning additive, and HB additive—with only 17 of the 32 most popular profiles (Numbers 16 and 17 were tied). The results show that DOCMP remained significantly better on the K-L percentages and best or not significantly different than best on hit rates, even though we are now estimating the models with approximately half the observations per respondent.

Until the issue of optimal DOC consideration experimental designs is resolved, the performance of DOC-based estimation methods remains a conservative predictive test. Improved or adaptive experimental designs might improve performance.

Summary of Empirical Results

The DOC-based estimation appears to predict hit rates well and to provide information (K-L percentage) about consideration decisions on validation data. Predictions appear to be better with DOC-based estimation than with any of the other five decision rules for both machine-learning and HB estimation, though an unconstrained machine-learning additive model (which can represent some noncompensatory rules) comes close. Some of this improvement is due to cognitive simplicity.

TARGET POPULATION, TASK FORMAT, AND PROFILE REPRESENTATION

We examine hypotheses that the predictive ability is unique to the task format, to the GfK respondents, or to the way we present profiles. We discuss each in turn.

Variations in Task Formats

With the format analyzed in the previous section, respondents must evaluate every profile (“evaluate all profiles”). However, such a restriction may be neither necessary nor descriptive. For example, Ordóñez, Benson, and Beach (1999) argue that consumers screen products by rejecting products they would not consider further. Because choice rules are context dependent (e.g., Payne, Bettman, and Johnson 1993), the task format could influence the propensity to use a DOC rule.

To examine context sensitivity, we tested alternative task formats. One format asked respondents to indicate only the profiles they would consider (“consider only”); another asked respondents to indicate only the profiles they would reject (“reject only”). The tasks were otherwise identical to “evaluate all profiles.” We also tested a “no-browsing” format, in which respondents evaluated profiles sequentially (in a randomized order). Representative screenshots for these formats, as well as feature introduction screenshots and instruction screenshots, are available in the Web Appendix (http://www.marketingpower.com/jmrjune10).

The predictive results mimic the results in Table 1.8 On the K-L percentages, both DOC-based estimation methods were significantly better than all benchmarks on all four for-

---

7We note the poor performance of the machine-learning subset conjunctive model with $S = 16$. With $S = 16$ and a goal of choosing 0 versus 1 for $w_{qij}$, the subset conjunctive integer program tends to overfit the calibration data.

8Tables for the other formats are available in the Web Appendix (http://www.marketingpower.com/jmrjune10).
mats. On hit rate, at least one of the DOC-based estimation methods was best on all formats, significantly better than all benchmarks for the majority of the formats (three of four), and significantly better than nine of the ten benchmarks for the remaining format. On hit rate, the only estimation method that did not differ significantly from a DOC-based estimation method on that one format was the machine-learning additive model—a result similar to that which we observed in Table 1. To test DOC-based methods further, we merged the data from the four formats and compared DOCMP and LAD-DOC hit rates with the additive machine-learning method. When the hit rate data are merged, both DOCMP and LAD-DOC predict significantly better than the additive machine-learning method (t = 4.4, p < .01; t = 3.0, p < .01).

As the evaluation cost theory of consideration-set formation predicts, respondents considered fewer profiles when the relative evaluation cost (for consideration) was higher: 4.3 profiles in “consider only,” 7.8 in “evaluate all,” and 10.6 in “reject only.” As the theory of context dependence predicts, the propensity to use a second DOC pattern varied as well. In addition, disjunctions were more common when consideration sets were larger: 0% for “consider only,” 7.1% for “evaluate all,” and 9.8% for “reject only.” Although the data cannot distinguish whether these differences are due to the size of the consideration set or due to differential evaluation costs induced by task variation, they illustrate how the predictive tests complement more direct (but possibly more intrusive) experimental measures.

U.S. Student Sample Versus Representative German Sample

We replicated the evaluate-all-profiles GPS measurement with a sample of MBA students at a U.S. university. Students were invited to an English-language Web site (e.g., first panel of Figure 1). As incentives, and to maintain incentive compatibility, they were entered in a lottery with a 1-in-25 chance of winning a laptop bag worth $100 and a 1-in-100 chance of winning a combination of cash and one of the GPSs that they considered. The response rate for U.S. students was lower (26%), and on average, consideration-set sizes were larger. Despite the differences in sample, response rate, incentives, and consideration-set size, DOCMP and LAD-DOC predicted validation data best (or were not significantly different than the best) on both hit rates and K-L percentages. (Again, the best benchmark was the additive machine-learning model. Lower sample sizes for the U.S. sample made it more difficult to distinguish differences.)

Text-Only Versus Visual Representation of the GPS Profiles

The profile representations in Figure 1 were designed by a professional graphic artist and were pretested extensively. Pretests suggested which features should be included in the JPEGs and which features should be included as satellite icons. Nonetheless, it is possible that the relative predictive ability of the estimation methods depends on the specific visual representations of the profiles. To examine this hypothesis, we included a task format that was identical to the task in “consider all profiles,” except that all features were described by text rather than pictures, icons, and text. Again, the DOC-based estimation methods are the best predictive methods—significantly better on K-L percentages and best or not significantly different than the best on hit rates—and again, the additive machine-learning method does as well on hit rate but not the K-L percentage. We cannot distinguish with the data whether this is a text-only effect or a result consistent with the analyses of the other formats. Notably, there is no significant difference in hit rates or K-L percentages between picture representations and text representations for either DOCMP or LAD-DOC.

Summary of Robustness Tests

The relative predictive ability of the tested methods appears to be robust with respect to the following:

- Format of the respondent task (evaluate all profiles, consideration only, rejection only, or no browsing),
- Respondent sample (representative German versus U.S. student), and
- Presentation of the stimuli (pictures versus text).

Managerial Implications and Diagnostics

We were motivated to study consideration-set decisions with a managerial challenge: How can a firm increase the likelihood that its products will be considered? We hope that by estimating DOC-based models, we might gain insight to help a firm enhance consideration. If the improved predictive ability of DOC-based models holds up to further testing, market-response simulators using DOC-based models might be more accurate than market-response simulators based on conjunctive, disjunctive, subset conjunctive, q-compensatory, or additive-rule decision rules (for a discussion of using predictive models to evaluate strategies, see Geisser 1993). To illustrate how models affect managerial decisions differently, we compare the simulated value of feature improvements between estimated DOC rules and estimated additive rules. (The data are available for readers who want to explore other comparisons.)

Changes in Market Share as a Function of Feature Improvements

Ofek and Srinivasan (2002, p. 401) propose that a value of a feature should be defined as “the incremental price the firm would charge per unit improvement in the product attribute (assumed to be infinitesimal) if it were to hold market share (or sales) constant.” In DOC rules, features and price levels are discrete; thus, we modify this definition slightly. We compute the incremental improvement in market share if a feature is added for an additional $50 in price. Because this calculation is sensitive to the base product, we select the features of the base product randomly. We illustrate two of the many differences between DOC rules and additive rules. In both situations, the recommended managerial decision depends on whether consumers consider products using the estimated DOC rules or the estimated additive rules.

Example 1

The DOC rules predict that consideration share will increase if we switch to the Garmin GPS and raise the price by $50, but compensatory rules predict that consideration share will decrease. To understand this difference intuitively, we recognize that the estimated DOC rules imply that 12% of the respondents screen on brand, and of those, 82% screen on Garmin. The remaining respondents screen on other features. With an additive partworth rule, 54% of the respondents have slightly higher partworths for
the Magellan GPS. With DOC rules, the advantage to Garmin comes from the 12% who screen on brand, but with additive rules, the advantage to Magellan comes a little from all the respondents in the sample.

Example 2. Additive rules predict that “extra bright” is the highest-valued feature improvement, yielding an 11% increase for the $50 price. However, the DOC rules predict a much smaller improvement (2%) because many of the respondents who screen on extra bright also eliminate GPSs with the higher price.

Diagnostic Summaries of DOC Rules

Diagnostic summaries of additive partworths have been developed through decades of application. Recent developments have added heterogeneity with corresponding challenges in how best to summarize heterogeneity to managers. Diagnostic summaries of noncompensatory decision rules are relatively nascent. Academics and practitioners are still evolving the best way to summarize such rules for managers.

This challenge is exacerbated for DOC rules. Even with cognitive simplicity (S = 4), there are 34,112 potential DOC patterns. Listing each pattern that matches consideration in a sample of respondents is not nearly as diagnostic as the feature-improvement simulations, which aggregate across identified patterns. As a first attempt, we examined summaries of first- and second-order inclusion. (Gilbride and Allenby [2004] and Yee and colleagues [2007] report first-order inclusion.) For example, the mini-USB port appeared in at least one DOC conjunction for 36% of the respondents. Extra-bright displays (25%) and color displays (21%) were the next-highest contributors. With second-order inclusions, for example, respondents who want a long battery life also want a mini-USB port (50%) and a color display (40%). Such first- and second-order conjunctive inclusions provide insight that complements DOC model-based market-response simulators. As in the market-response simulations, these simple diagnostics vary from what might be inferred from additive partworths.

We hope that such diagnostic information combined with market-response simulators will help managers evaluate product line changes and marketing activities. With more experience, researchers might develop more intuitive ways to summarize DOC patterns for managers.

SUMMARY AND FUTURE DIRECTIONS

Consideration sets have become relevant to managerial decisions in many product categories, and whenever there are many products available and/or products are described by many features and levels, extant research suggests that consumers use noncompensatory decision rules to make consideration decisions. Research further suggests that such decision rules are often cognitively simple. We hope we have contributed to these literature streams.

Drawing on qualitative research, we propose a generalization of established noncompensatory decision rules: DOC. We posit further that DOC rules will be cognitively simple and that models that attempt to represent cognitively simple DOC rules will predict better than models that do not. We examine two machine-learning estimation methods, DOCMP and LAD-DOC, and compare predictions with five decision-rule models as implemented by both machine-learning and HB estimation methods.

The preponderance of the empirical evidence in this article suggests that DOC rules and both estimation algorithms are worth further investigation. Both are significantly better on K-L percentages for all ten benchmarks, all four respondent task formats, German and U.S. data, and both highly visual and text-only stimuli. We obtain the same perspective with hit rates with one important exception. The machine-learning additive method does almost as well for some formats, a result that is consistent with the known robustness of the additive model and with ability of the additive model to represent some noncompensatory decision rules.

The results must be considered hypotheses for further testing. The handheld GPS category has many features, and at the time of testing, it was relatively new to the respondents. This provided a “proof-of-concept” test for DOC-based methods. In more familiar or simpler categories, additive models might suffice. Conversely, more complex categories, such as automobiles, might favor DOC rules.

We chose two methods to estimate DOC rules. There are likely others. For example, decision trees can also represent DOC rules (Breiman et al. 1984; Currim, Meyer, and Le 1988). If researchers can develop a way to model cognitive simplicity on decision trees, this approach might prove promising. If features are continuous, DOC rules are similar to specific interactions in a multilinear decision rule (Bordley and Kirkwood 2004; Mela and Lehmann 1995). With sufficient creativity and experimentation, researchers might extend finite-mixture, Bayesian, simulated maximum likelihood, Markov, or kernel estimators to estimate cognitively simple continuous DOC analogs (Evgeniou, Boussios, and Zacharia 2005; Hauser and Wisniewski 1982; Mela and Lehmann 1995; Rossi and Allenby 2003; Swait and Erdem 2007).

Finally, we focused on the consideration stage of a consider-then-choose rule. The DOC rules might also apply to the choice stage. A model that is DOC in the first stage and compensatory in the second stage might also be investigated. There is a rich history in marketing of two-stage models in which consideration is a latent, unobserved construct (e.g., Andrews and Srinivasan 1995; Gensch 1987; Gilbride and Allenby 2004; Siddharth, Bucklin, and Morrison 1995; Swait and Erdem 2007). We believe that DOC rules combined with cognitive simplicity could complement these lines of research.

REFERENCES


