PSYCHOMETRIKA—VOL. 70, NO. 4, 737–757
DECEMBER 2005
DOI: 10.1007/s11336-002-0978-8

PROBABILISTIC SUBSET CONJUNCTION

RAJEEV KOHLI AND KAMEL JEDIDI

GRADUATE SCHOOL OF BUSINESS, COLUMBIA UNIVERSITY

The authors introduce subset conjunction as a classification rule by which an acceptable alternative must satisfy some minimum number of criteria. The rule subsumes conjunctive and disjunctive decision strategies as special cases.

Subset conjunction can be represented in a binary-response model, for example, in a logistic regression, using only main effects or only interaction effects. This results in a confounding of the main and interaction effects when there is little or no response error. With greater response error, a logistic regression, even if it gives a good fit to data, can produce parameter estimates that do not reflect the underlying decision process. The authors propose a model in which the binary classification of alternatives into acceptable/unacceptable categories is based on a probabilistic implementation of a subset-conjunctive process. The satisfaction of decision criteria biases the odds toward one outcome or the other. The authors then describe a two-stage choice model in which a (possibly large) set of alternatives is first reduced using a subset-conjunctive rule, after which an alternative is selected from this reduced set of items. They describe methods for estimating the unobserved consideration probabilities from classification and choice data, and illustrate the use of the models for cancer diagnosis and consumer choice. They report the results of simulations investigating estimation accuracy, incidence of local optima, and model fit.

Key words: noncompensatory decision strategies, conjunctive/disjunctive strategies, subset-conjunctive strategies, consideration sets, binary data, choice models, maximum likelihood estimation.

1. Introduction

First brought to notice by H.M. Johnson in 1935 (Coombs, 1951), conjunctive and disjunctive decision rules have been widely studied in psychology, and related disciplines like marketing (e.g., Teigen, Martinussen & Lund, 1996; Westenberg & Koele, 1994; Grether & Wilde, 1984; Lussier & Olshavsky, 1979; Payne, 1976; Payne, Bettman & Johnson, 1988; Wright & Barbour, 1977; Wright, 1975). The two rules are used for screening large sets of alternatives. Examples are the screening of applicants (e.g., for jobs by employers and mortgage applications by banks; see Dawes 1979 and Grether & Wilde, 1984) and the formation of consideration- or choice-sets by consumers (e.g., Huber & Klein, 1991; Roberts & Lattin, 1991; Andrews & Srinivasan, 1995). Marketing managers often use conjunctive rules when defining target markets over demographic variables; and doctors use such rules when diagnosing illnesses based on a conjunction of symptoms.

Maris (1999) describes the use of conjunctive and disjunctive rules for representing cognitive processes in latent response models. Einhorn (1970) considers a mathematical model that approximates conjunctive and disjunctive rules for continuous predictor variables. Mela & Lehmann (1995) propose a method for inferring conjunctive and disjunctive rules in a regression framework. Boros, Hammer & Hooker (1994, 1995), Van Mechelen (1988), and Leenen & Van Mechelen (1998) examine Boolean regression for inferring the rules from binary response data. Swait (2001) considers a penalized utility function in a compensatory utility-maximization

The authors thank the Editor, the Associate Editor, and three anonymous reviewers for their constructive suggestions, and also thank Asim Ansari and Raghuram Iyengar for their helpful comments. They also thank Sawtooth Software, McKinsey and Company, and Intelliquest for providing the PC choice data, and the University of Wisconsin for making the breast-cancer data available at the machine learning archives.

Request for reprints should be sent to Kamel Jedidi, Columbia University, Graduate School of Business, 518 Uris Hall, 3022 Broadway, New York, NY 10027, USA. E-mail: kj7@columbia.edu.

framework that allows representation of conjunctive and disjunctive choice strategies and their combinations.

We introduce subset conjunction, a screening rule by which an acceptable alternative satisfies a minimum number of criteria, not necessarily one criterion (disjunctive) or all criteria (conjunctive). A person who favors a conjunctive strategy, but finds few acceptable alternatives, has two options: either accept some previously unacceptable attribute levels; or reduce the number of criteria, say t, that an acceptable alternative must satisfy. Similarly, if there are too many acceptable alternatives, a person who otherwise uses a disjunctive strategy can increase the value of t, or can switch acceptable attribute levels to unacceptable. In this sense, a subset conjunctive rule offers flexibility, allowing one to vary the size of an acceptable subset by changing the number of criteria satisfied by an acceptable alternative.

The proposed decision strategy subsumes conjunctive and disjunctive rules as special cases. In principle, it is a special case of Barthelemy & Mullet's (1987) algebraic model of categorical judgment (see also Montgomery, 1983; Barthelemy & Mullet, 1996) and of Crama, Hammer & Ibaraki's (1988) partially defined Boolean functions. We restrict attention to discrete attributes and show that subset-conjunctive rules can be represented by binary response models that contain only main effects, or only interaction effects. Consequently, the main effects are confounded in these models with the interaction effects if there is little or no error in the response variable. With more error in responses, a logistic regression produces good overall fits to the data but finds parameter estimates that do not reflect the underlying decision process. In this paper, we propose a probabilistic model that generalizes the deterministic form of a subset-conjunctive rule. Each attribute level is acceptable not with certainty, but with a probability, which can be unobserved. We then describe an extension of the model to choice data with consideration modeled as an unobserved step preceding the choice of an alternative. We describe methods for estimating the models from binary and multinomial choice data, and illustrate them with examples from cancer diagnosis and consumer psychology. Finally, we report simulation results relating to estimation accuracy, model fit, and incidence of local optima.

2. Binary Response

Let m denote the number of attributes or decision criteria. Let attribute k have n_k discrete (nominal or ordinal) levels, $1 \leq k \leq m$. An alternative has one level of each attribute. Thus, there are at most $\prod_{k=1}^{m} n_k$ possible alternatives. A person's evaluation is a classification of an alternative into one of $R \geq 2$ categories. For simplicity, we consider $R = 2$, in which case a person classifies an alternative i as acceptable if it is satisfactory on at least t attributes, $1 \leq t \leq m$; otherwise, the person classifies the alternative as unacceptable. The condition $t = 1$ gives a disjunctive rule; and the condition $t = m$ gives a conjunctive rule.

Suppose there is no response error. We can represent a subset-conjunctive strategy by a linear model in the following manner. Let $x_k \in \{0, 1\}$ represent the kth of m binary attributes; let $x_k = 1(0)$ denote an acceptable (unacceptable) attribute level, $1 \leq k \leq m$. Consider $m = 2$. Suppose a person evaluates an alternative to be acceptable if it is acceptable on both attributes; otherwise s/he finds the alternative to be unacceptable. That is,

$$y = \begin{cases} 1, & \text{if } x_1 = x_2 = 1, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $y = f(x_1, x_2) = 1(0)$ if an alternative is acceptable (unacceptable). As x_1 and x_2 are 0–1 variables, we can replace $x_1 = x_2 = 1$ in (1) by the condition $x_1 x_2 \geq 1$; i.e.,

$$y = \begin{cases} 1, & \text{if } -1 + x_1 x_2 \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Equation (2) has the familiar form of a binary logistic/probit model with an interaction term, albeit without error.[1]

Now suppose a person finds an alternative acceptable if it is acceptable on at least $t = 2$ of the $m = 3$ possible binary attributes; i.e.,

$$y = \begin{cases} 1, & \text{if } (x_1 = x_2 = 1) \text{ or } (x_1 = x_3 = 1) \text{ or } (x_2 = x_3 = 1), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We replace $x_k = x_\ell = 1$ by $x_k x_\ell \geq 1$ in (3), $1 \leq k < \ell \leq 3$. Then

$$y = \begin{cases} 1, & \text{if } (x_1 x_2 \geq 1) \ \vee (x_1 x_3 \geq 1) \ \vee (x_2 x_3 \geq 1), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The disjunctive "or" ($\vee$) in the above expression can be represented by the condition that the sum of the two-way products exceeds zero; i.e.,

$$y = \begin{cases} 1, & \text{if } -1 + x_1 x_2 + x_1 x_3 + x_2 x_3 \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Equation (5) has the form of a binary logistic/probit model with all $t = 2$ way interaction effects among the $m = 3$ binary variables.

In general, an error-free binary-response model with all t-factor interaction effects and an intercept ($= -1$) describes a subset-conjunctive strategy requiring an acceptable alternative to satisfy at least t criteria, $2 \leq t \leq m$.[2] That is, we can represent

$$y = \begin{cases} 1, & \text{if } x_k = 1 \text{ for at least } t \text{ values of } 1 \leq k \leq m, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

by the interaction-effects model

$$y = \begin{cases} 1, & \text{if } -1 + \sum_{\alpha \in A_t} x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3} \ldots x_m^{\alpha_m} \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where

$$A_t = \left\{ \alpha \Big| \sum_{k=1}^{m} \alpha_k = t \right\}, \qquad 1 \leq t \leq m, \tag{8}$$

and

$$\alpha = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_m\}, \qquad \alpha_k \in \{0, 1\}, \quad x_k^0 \equiv 1, \quad x_k \in \{0, 1\}, \quad 1 \leq k \leq m. \tag{9}$$

There is another way of representing a subset-conjunctive strategy in a linear model. As the x_k are 0–1 variables, the condition $x_1 + x_2 + x_3 + \cdots + x_m \geq t$ is satisfied only if at least t of the m values of x_k are ones. Thus,

$$y = \begin{cases} 1, & \text{if } -t + \sum_{k=1}^{m} x_k \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

implies that an alternative is acceptable ($y = 1$) if it is acceptable on at least t of m attributes (i.e., if at least t of the m x_k's are ones). For example,

$$y = \begin{cases} 1, & \text{if } -2 + x_1 + x_2 \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

[1]The corresponding logistic regression is

$$y = \begin{cases} 1, & \text{if } \beta_0 + \beta_1 x_1 x_2 + \epsilon \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where in this case $\beta_0 = -1$, $\beta_1 = 1$, and ϵ is a logistic error term.

[2]For $t = 1$, the corresponding representation has only main effects and an intercept term with value -1.

represents a conjunction over two binary attributes, and

$$y = \begin{cases} 1, & \text{if } -2 + x_1 + x_2 + x_3 \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

represents a subset conjunction in which an acceptable alternative is satisfactory on at least two of the three attributes.

In summary, an error-free subset-conjunctive process can be represented in at least two ways by the model

$$y = \begin{cases} 1, & \text{if } \beta_0 + \sum_k \beta_k x_k + \sum_{l<k} \beta_{lk} x_l x_k + \sum_{j<k<\ell} \beta_{jk\ell} x_j x_k x_\ell + \cdots \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where the β's are the model coefficients. One representation has only interaction effects of order t. The other representation has only main-effects terms. A model with both sets of terms cannot be estimated because the main effects are confounded with the interaction effects.[3]

One can use this result to infer a subset-conjunctive process from (say) a logistic regression when there is low response error in the data. In a main-effects model:

1. the coefficient estimates should all have nearly the same value, $\beta_k \approx \beta$;
2. the intercept should be $\beta_0 \approx -t\beta$, $1 \leq t \leq m$; and
3. the logit for all items should fall in the extreme tails of the distribution so that all probabilities are close to zero or one; i.e., $\beta_0 + (t-1)\beta \ll 0$ and $\beta_0 + t\beta \gg 0$.

But this method of inferring a subset-conjunctive rule fails if the response error is large. For example, we find in two simulated examples with larger response error ($m = 5, t = 2$) that the main and interaction effects are not confounded, but it is also not possible to interpret the underlying process as a subset conjunction: a stepwise logistic regression finds all significant main effects, a few significant second-order interaction effects, and a significant third-order interaction effect with a negative coefficient.[4]

How then should one infer a subset-conjunctive rule from data? If there is no response error, a deterministic conjunctive (disjunctive) rule can be identified by setting an attribute level to be acceptable (unacceptable) only if it appears in at least one acceptable (unacceptable) alternative. For data with error, Boros et al. (1994, 1995), Van Mechelen (1988), and Leenen and Van Mechelen (1998) examine deterministic algorithms for this Boolean regression problem. We examine a probabilistic formulation that allows for binary and multinomial choice responses and permits any subset-conjunctive process.

2.1. Probabilistic Subset Conjunction

Let π_{jk} denote the probability with which level j of attribute k is acceptable, $1 \leq j \leq n_k$, $1 \leq k \leq m$. Deterministic conjunctive/disjunctive models assign 0-1 values to the probabilities. We assume that the acceptability of one attribute level is independent of the acceptability of another attribute level.

[3]Observe that this is a different source of confounding than the more familiar variety arising from the aliasing of main and interaction effects in an experimental design (which in a binary response model affects the estimation of the unobserved response function). One can have data from a full factorial design, but the parameter estimates will continue to be confounded if the data arise from a subset-conjunctive process.

[4]Logistic regressions with only main effects, or with only two-way interaction effects, find all significant parameter estimates and comparable likelihood values. Neither pattern of parameter estimates suggests that the underlying process is a subset conjunction. That regression can produce good model fits, without reflecting an underlying decision process, is well known; see, e.g., Dawes and Corrigan (1974).

Let $x_{ijk} = 1(0)$ if level j of attribute k appears (does not appear) in alternative i. Then the probability that alternative i is acceptable on attribute k is

$$p_{ik} = \prod_{j=1}^{n_k} \pi_{jk}^{x_{ijk}}. \tag{14}$$

Let $q_{ik} = 1 - p_{ik}$. Let $\theta_i^s(k, m)$ denote the probability that alternative i is acceptable on at least s of the $(m - k + 1)$ attributes $k, k + 1, k + 2, \ldots, m$. For example,

$$\theta_i^1(k, m) = 1 - \prod_{u=k}^{m} \prod_{j=1}^{n_u} (1 - \pi_{ju})^{x_{iju}} \tag{15}$$

denotes the probability that alternative i is acceptable on at least one of the attributes k through m. Let

$$\theta_i^s(k, m) = 0 \qquad \text{for} \quad \text{all } s > m - k + 1. \tag{16}$$

Then the probability that alternative i is acceptable on at least t of m attributes is given by the recursion

$$\pi_i^t = \theta_i^t(1, m) = p_{i1}\theta_i^{t-1}(2, m) + q_{i1}\theta_i^t(2, m). \tag{17}$$

The disjunctive ($t = 1$) and conjunctive ($t = m$) rules have the particularly simple forms

$$\pi_i^1 = 1 - \prod_{k=1}^{m} q_{ik} = 1 - \prod_{k=1}^{m} \prod_{j=1}^{n_k} (1 - \pi_{jk})^{x_{ijk}}, \qquad \pi_i^m = \prod_{k=1}^{m} p_{ik} = \prod_{k=1}^{m} \prod_{j=1}^{n_k} \pi_{jk}^{x_{ijk}}. \tag{18}$$

The recursion can be used to write π_i^t in terms of the π_{jk} for all $2 \le t \le m - 1$. For example,

$$\pi_i^2 = p_{i1}\theta_i^1(2, m) + q_{i1}\theta_i^2(2, m) \tag{19}$$

is the probability that alternative i is acceptable on at least $t = 2$ attributes. The right-hand side can be expanded by successively substituting

$$\theta_i^1(k, m) = p_{ik} + q_{ik}\theta_i^1(k + 1, m), \tag{20}$$

$$\theta_i^2(k, m) = p_{ik}\theta_i^1(k + 1, m) + q_{ik}\theta_i^2(k + 1, m), \qquad 2 \le k \le m. \tag{21}$$

As an illustration, consider a problem with $m = 7$ attributes, each with $n_k = 3$ levels. Let

$$\theta_i^1(k, 7) = 1 - \prod_{\ell=k}^{7} \prod_{j=1}^{3} (1 - \pi_{j\ell})^{x_{ij\ell}}. \tag{22}$$

Then the probability that alternative i is acceptable on at least $t = 2$ attributes is given by

$$\begin{aligned} \pi_i^2 = {} & p_{i1}\theta_i^1(2, 7) + \ q_{i1}\big(p_{i2}\theta_i^1(3, 7) + \ q_{i2}\big(p_{i3}\theta_i^1(4, 7) + \ q_{i3}\big(p_{i4}\theta_i^1(5, 7) \\ & + q_{i4}\big(p_{i5}\theta_i^1(6, 7) + \ q_{i5}p_{i6}p_{i7}\big)\big)\big)\big). \end{aligned} \tag{23}$$

Let N_h denote the number of alternatives evaluated by subject h, where $1 \le h \le N$. Let y_{ih} be a dummy variable that indicates whether subject h finds alternative i acceptable (1) or unacceptable (0), $1 \le i \le N_h$.[5] Estimates of π_{jk} can be obtained by maximizing the likelihood

[5]The set of alternatives evaluated by a subject can be the same or different across subjects. Technically, we should have used subscript i_h to denote the ith alternative evaluated by subject h. We avoid such notation, however, for simplicity.

function

$$L_t = \prod_{h=1}^{N} \prod_{i=1}^{N_h} \left(\pi_i^t\right)^{y_{ih}} \times \left(1 - \pi_i^t\right)^{1-y_{ih}}, \tag{24}$$

where $0 \leq \pi_{jk} \leq 1$, for all $1 \leq j \leq n_k$, $1 \leq k \leq m$. Additionally, one can restrict the estimates to satisfy a priori preference orderings on the attribute levels. The likelihood function can be maximized using standard nonlinear optimization packages (e.g., PROC NLP in SAS) for problems with under 30 predictor dummies. Larger problems require specialized algorithms, two of which are described in the Appendix.

As the π_i^t are probabilities, the likelihood function is naturally scaled between zero and one. The upper bound on the likelihood (log-likelihood) value is one (zero), which happens when all acceptable alternatives have $\pi_i^t = 1$ and all unacceptable alternatives have $\pi_i^t = 0$. The least informative model corresponds to the case where $\pi_i^t = \pi^t$.

Let $N_A = \sum_{h=1}^{N} \sum_{i=1}^{N_h} y_{ih}$ denote the number of acceptable alternatives. Let $N_U = \sum_{h=1}^{N} \sum_{i=1}^{N_h} (1 - y_{ih})$ denote the number of unacceptable alternatives. Then a naive estimate for the acceptance probability of each alternative is obtained by setting $\hat{\pi}^t = N_A/(N_A + N_U)$. The ratio of the estimated likelihood function L_t to its value $L(0)$ at $\hat{\pi}^t$ gives the relative odds for the two models. A measure of the improvement in the log-likelihood value relative to that of the naive model is $\rho^2 = 1 - [\log L_t / \log L(0)]$ (see Ben-Akiva & Lerman, 1993, p. 167).

An attribute k is irrelevant if all levels j have a common consideration probability π_{jk}. Such an attribute can be eliminated from the model if the consideration probabilities are all zeros or all ones, the value of t being reduced by one in the latter case. A good model should have at least some π_{jk} biased toward the extreme values, and there should be within-attribute differences in the π_{jk} values.

2.2. *Example*

Fine needle aspiration (FNA) is a method for extracting a sample of cells from a patient. It is used in breast cancer diagnosis, after a patient has developed a lump in a breast. The nuclei of the extracted cells are examined for abnormal characteristics and used to assess if a patient has breast cancer.

We examine data from 569 patients, 212 with malignant tumors and 357 with benign tumors, to see if the presence of breast cancer can be predicted by a subset-conjunctive rule defined over seven cell characteristics listed in the leftmost column of Table 1. The seven characteristics are taken from a larger set of 30 variables that Wolberg, Street, Heisey & Mangasarian (1995) use in a linear-programming model for classifying tumors into malignant and benign categories.[6]

Table 1 shows the parameter estimates for disjunctive, conjunctive, and five subset-conjunctive models, with subset sizes varying from two to six.[7] The lowest log-likelihood value corresponds to a subset-conjunctive model requiring at least $t = 2$ of the seven criteria to be "satisfied" for a malignant classification. With a $\rho^2 = 0.775$, this model is a significant improvement over the naive model. The estimated probabilities (Table 1, $t = 2$) suggest the following conclusions:

[6] As the present model only allows discrete characteristics, we construct three categories for each of the seven originally continuous variables, using the 25th percentile and the 75th percentile of the observed values as cutoff points.

[7] Parameter estimates are obtained by maximizing the likelihood function in (24) for $t = 1, \ldots, 7$, where the probability that alternative i is acceptable on at least t of the $m = 7$, three-level attributes, π_i^t, is derived recursively using equation (17). Equation (23) provides an example for the special case $t = 2$.

TABLE 1.
Breast cancer diagnosis using the probabilistic subset-conjunction rule.

Variable	Subset size (t)						
	Conjunctive 7	6	5	4	3	2	Disjunctive 1
Std. error of radius							
x_{11} Low	0.43	0.77	0.77	0.00	0.00	0.00	1.00
x_{21} Medium	1.00	1.00	1.00	0.13	0.09	0.00	1.00
x_{31} High	1.00	1.00	1.00	1.00	1.00	0.27	0.79
Std. error of compactness							
x_{12} Low	1.00	1.00	0.00	0.00	0.00	0.00	1.00
x_{22} Medium	1.00	1.00	0.00	0.00	0.00	0.00	0.95
x_{32} High	1.00	1.00	0.00	0.00	0.00	0.00	1.00
Worst radius							
x_{13} Low	0.04	0.00	0.00	0.00	0.00	0.00	1.00
x_{23} Medium	0.67	0.05	0.05	0.34	0.31	0.06	1.00
x_{33} High	1.00	1.00	1.00	1.00	1.00	1.00	0.03
Worst texture							
x_{14} Low	0.48	0.19	0.19	0.00	0.00	0.00	1.00
x_{24} Medium	1.00	0.84	0.84	0.23	0.31	0.00	1.00
x_{34} High	1.00	1.00	1.00	1.00	1.00	0.34	0.75
Worst smoothness							
x_{15} Low	1.00	0.92	0.92	0.00	0.00	0.00	1.00
x_{25} Medium	1.00	1.00	1.00	0.00	0.00	0.00	1.00
x_{35} High	1.00	1.00	1.00	1.00	0.35	0.05	0.94
Worst concavity							
x_{16} Low	0.27	0.00	0.00	0.00	0.00	0.08	1.00
x_{26} Medium	1.00	1.00	1.00	1.00	1.00	1.00	1.00
x_{36} High	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Worst # concave points							
x_{17} Low	0.06	0.00	0.00	0.00	0.00	0.00	1.00
x_{27} Medium	0.66	0.18	0.18	0.00	0.00	0.00	1.00
x_{37} High	1.00	1.00	1.00	1.00	1.00	0.93	0.05
Minus log-likelihood	134.13	97.22	97.22	93.13	85.19	84.38	96.01

Subset size $t = 2$ is best solution.

1. Malignancy is mainly determined by three characteristics: high "worst radius," medium-to-high "worst concavity," and high "worst number of concave points."
2. Tumors with two of the above three characteristics are almost certainly malignant.
3. Tumors with one of the above three characteristics have a moderate risk of malignancy if the "standard error of radius" and/or the "worst texture" have a high value.
4. Tumors are benign if they have:
   (a). low values of "worst radius" and "worst concavity"; and
   (b). low-to-medium values of "standard error of radius," "worst radius," "worst texture," "worst smoothness," and "worst number of concave points."
5. All other tumors carry a low risk for breast cancer.

A stepwise logistic regression, with selection entry set at the $p = .05$ level, also fits the data well. The best-fitting model has only main-effects terms; the estimated logistic regression

equation is (see Table 1, column 1, for variable definitions):

$$u = 10.05 - 8.64x_{13} - 5.43x_{23} - 3.53x_{14} - 1.97x_{24} - 7.82x_{17} - 4.9x_{27}, \qquad LL = -83.84,$$

where u is the logit of the probability of accepting the alternative and LL is the log-likelihood value. The logistic-regression and subset-conjunctive models are similar in terms of their overall fits, but can make different predictions because the former has a compensatory structure and the latter has a noncompensatory structure. For example, setting $x_{13} = x_{17} = 1$ and $x_{23} = x_{14} = x_{24} = x_{27} = 0$ in the logistic regression gives

$$u = 10.05 - 8.64 - 7.82 = -6.41, \quad p(\text{malignancy}) = \frac{e^{-6.41}}{1 + e^{-6.41}} = 0.0016.$$

That is, cell samples with "low" worst radius ($x_{13} = 1$), "low" worst number of concave points ($x_{17} = 1$), and "high" values on the other characteristics, are associated with a very small probability of malignancy. The corresponding subset-conjunctive model with $t = 2$ gives a probability of malignancy that exceeds $0.27 + 0.34 = 0.61$ (see Table 1). Unfortunately, there are no cases in the sample to test for this difference (and other similar differences) in the predictions of the two models, for the reason that cancerous cells are simultaneously altered on several cell features. We therefore cannot say that the underlying process is a subset conjunction, but only that the data are consistent with the process, and that there are conditions (albeit unobserved in the present instance) where the outcomes can differ substantially from the predictions of a logistic regression.

To use the subset-conjunctive model for cancer diagnosis, one has to select a probability cutoff for classifying a tumor as benign or malignant. Although the value of this cutoff should depend on the relative importance of the type I and type II errors (it is evidently more costly to misdiagnose a malignant tumor), we ignore this here and assume that the two types of errors are equally important. That is, we classify a tumor as malignant if $\pi_i^t > 0.50$. The corresponding hit rate is 94.20%, which is identical to that obtained from logistic regression but much higher than the 62.74% predicted by the maximum-chance criterion.

A test of predictive validity is obtained by running a ten-fold cross validation. We randomly select 90% of the observations for model estimation and use the remaining 10% for prediction. We repeat this analysis 100 times. The mean hit rate across randomly drawn, holdout samples is 94.53%, and the 95% confidence interval is [89.01%, 98.43%]. The first of these is slightly lower than the hit rate obtained using a linear-programming classification (Wolberg et al., 1995), which has a mean value of 97.5% in ten-fold cross validations. We suspect two reasons for the poorer performance: the use of seven instead of thirty predictor variables, and the need to discretize these variables to use our model. For comparison, the mean hit rate from logistic regression is 93.53% and the 95% confidence interval is [87.71%, 98.01%]. Both measures are slightly worse than those obtained from the subset-conjunctive model.

Computationally, the maximum likelihood procedure performs reasonably well in this example. Fifty two of the 100 runs with random starts converged to the maximum likelihood value of -84.38; 21 runs converged to a value of -94.11; 12 runs converged to a value of -95.36; and 15 runs converged to a value of -98.59. Although the proportion of runs that converged to the largest maximum-likelihood value is reasonably high, this result emphasizes the need to rerun the estimation procedure several times (e.g., ≥ 10 times) in any application to ensure that proper convergence is achieved. This task can be performed quickly since the procedure's convergence time is generally low (an average of 43 seconds for this application). An alternative approach is to use rational starting values (see the Appendix). Finally, and more importantly, the parameter estimates across the 52 converged runs are identical. This suggests that the subset-conjunctive model is identified and does not suffer from model equivalence problems. Section 4 reports the results of a simulation study that investigates the incidence of local optima in subset-conjunctive models.

3. The Multinomial Subset-Conjunctive Model

The extension to multinomial response is useful when one has reason to believe that a subset-conjunctive strategy underlies the choice among alternatives in a choice set. For example, consumer choice is often described as a phased decision process in which consumers first screen options for further consideration and then make a choice among the considered alternatives (Payne et al., 1988). Other examples are physicians prescribing one of several possible drugs, and parents choosing one of several acceptable pre-schools for their children. We extend the above (binary) subset-conjunctive model to these situations, inferring from choice data the unobserved consideration rule used by a person. To achieve this, we have to extend the formulation to predict how a person chooses among screened options. Consideration becomes a *latent* variable, which we have to infer from choice data; and choice becomes conditional on consideration, each consideration set being assigned a probability that depends on the subset-conjunctive rule and the consideration probabilities assigned to the attribute levels.

Consider a choice set S, with $|S| = R$ alternatives. As in the binary case, let π_i^t denote the probability that alternative i is acceptable to a person using a subset-conjunctive rule. Then a person chooses none of the R alternatives if they are all unacceptable. In this case, the *no-choice* probability is

$$P^t(\bar{S}) = \prod_{i=1}^{R} \left(1 - \pi_i^t\right). \tag{25}$$

If the person makes a choice, then at least one of the R alternatives is considered. The alternatives in choice set S define 2^R possible subsets, including the empty (no choice) set. Let ω denote a particular subset and let $\Omega = \{\omega\}$ denote the set of all subsets, $|\Omega| = 2^R$. Then the probability that the *consideration set* is ω (i.e., that only alternatives in ω are acceptable) is

$$\pi^t(\omega) = \prod_{i \in \omega} \pi_i^t \prod_{i \in \mathbf{S} \setminus \omega} \left(1 - \pi_i^t\right). \tag{26}$$

Given a consideration set $\omega \in \Omega$, we wish to specify the *conditional* probability of choice $P_i^t(\omega)$ for each alternative $i \in \omega$. Let $P_i^t(\omega)$ be an increasing function of the consideration probabilities; i.e.,

$$P_i^t(\omega) = f\left(\pi_i^t | i \in \omega\right), \quad \omega \in \Omega \tag{27}$$

where

$$0 \leq P_i^t(\omega) \leq 1, \quad \sum_{i \in \omega} P_i^t(\omega) = 1, \quad \frac{\partial f}{\partial \pi_i^t} > 0, \quad i \in \omega. \tag{28}$$

A particularly simple functional form is obtained by imposing the proportionality condition

$$P_i^t(\omega) \propto \pi_i^t. \tag{29}$$

In this case, the conditional probability of choice given consideration set ω is

$$P_i^t(\omega) = \frac{\pi_i^t}{\sum_{\ell \in \omega} \pi_\ell^t}, \qquad i \in \omega. \tag{30}$$

One can use alternative or modified forms for the function $f(\cdot)$ — a priori, there is no basis for choosing one over another. We will use the above form because it is simple, relates to the Luce (1959) choice axiom, and has the form of the so-called "attraction models" in marketing (Cooper, 1993). The unconditional choice probability for item i in choice set S is thus

given by

$$P_i^t(S) = \sum_{\omega \in \Omega} P_i^t(\omega) \cdot \pi^t(\omega), \quad P_i^t(\omega) \equiv 0 \quad \text{if} \quad i \notin \omega. \tag{31}$$

Let y_{ih} indicate whether (1) or not (0) person h selects alternative i from choice set S, $1 \leq h \leq \mathrm{N}$. Let Ψ_h denote the set of choice sets evaluated by person h $1 \leq h \leq N$.[8] Given a random sample of N homogeneous subjects, estimates for the values of the consideration probabilities π_{jk} and the optimal subset-size t, $1 \leq t \leq m$, are obtained by maximizing the likelihood function

$$L = \prod_{h=1}^{N} \prod_{S \in \Psi_h} \left([P^t(\bar{S})]^{(1-\sum_i y_{ih})} \prod_{i \in S} [P_i^t(S)]^{y_{ih}} \right), \tag{32}$$

where $\sum_i y_{ih} = 1(0)$ if a person makes a (no) choice.[9] We maximize the function for each value of $1 \leq t \leq m$ and select the value of t for which the likelihood function has the largest value.

To illustrate the computation of the elements of the likelihood function, consider a conjunctive rule with m attributes ($t = m$), and a choice set S of size $R = 2$. There are three nonempty consideration subsets in this case: $\omega_1 = \{1\}$ and $\omega_2 = \{2\}$, which contain alternatives 1 and 2, respectively, and $\omega_3 = \{1, 2\}$, which contains both items 1 and 2. Let

$$\pi_i^m = \prod_{k=1}^{m} \prod_{j=1}^{n_k} \pi_{jk}^{x_{ijk}} \tag{33}$$

denote the probability that alternative i is acceptable to a person using a conjunctive rule. Then it is easy to verify that

$$P_i^m(S) = 1 \times \pi_i^m \times \left(1 - \pi_l^m\right) + \frac{\pi_i^m}{\left(\pi_i^m + \pi_l^m\right)} \times \left(\pi_i^m \times \pi_l^m\right), \quad i, l = 1, 2, \tag{34}$$

and

$$P^m(\bar{S}) = \left(1 - \pi_1^m\right) \times \left(1 - \pi_2^m\right). \tag{35}$$

It is important to note that the parameters π_{jk} are uniquely identified only if the data collection allows a "no-choice" option for each choice set. Intuitively, it is easy to see that without a "no-choice" option, it is not possible to observe when an alternative is unacceptable. Technically, the proportionality relation between choice and consideration probabilities implies that in the absence of a no-choice option, the consideration probabilities can be estimated up to a proportionality constant. The inclusion of a no-choice option eliminates this indeterminacy, because the "no-choice" probability is equal (not just proportional) to the probability that no item is considered in a choice set (see equation (25)).

3.1. A Consumer Psychology Example

We examine data obtained from 326 people participating in a commercial study of personal computer preferences. Every person was presented a sequence of eight independent choice sets, which differed across respondents and were selected from a master experimental plan. A person either rejected all alternatives in a choice set or selected one that was his/her most preferred. All choice sets were constructed to have three alternatives, each of which was characterized using

[8] The choice sets Ψ_h can be the same or vary across subjects. Technically, we should have used subscript S_h to denote the choice set evaluated by subject h. We avoid such notation, however, for simplicity.

[9] It is important to note here that the index h is used to denote observations and not parameters. This is because we are assuming homogeneous subjects. Ideally, the model should account for preference heterogeneity across subjects in which case the parameters need to be indexed by h as well. The authors are presently working on such an extension.

five attributes:

1. Brand (A, B, C, D, E);
2. Performance (below average, average, above average);
3. Warranty period (90 days, 1 year, 5 years);
4. Service location (ship back to manufacturer for service, service at local dealer, on-site service); and
5. Price (Low, Med–Low, Med–High, High) .[10]

Ideally, the analysis of such choice data should account for preference heterogeneity, which in the proposed multinomial model lies in the form of a subset-conjunctive rule (reflected in the value of t) and in the values of the consideration probabilities. The required extensions—in Bayesian or finite-mixture frameworks—are nontrivial and are currently being developed. Here, we restrict ourselves to estimating aggregate multinomial subset-conjunctive models and comparing these with the following nested multinomial logit model (see Maddala, 1983, pp. 67–70).

Let U_{ih} denote the latent utility person h has for choice alternative $i \in S$, where each choice set $S \in \Psi_h$ has three alternatives. Following the tradition in random-utility models (e.g., McFadden, 1973), we write $U_{ih} = V_{ih} + e_{ih}$, where e_{ih} is an error term that follows an extreme value distribution and V_{ih} is a linear function of the observed product attributes; i.e.,

$$V_{ih} = \sum_{j=2}^{5} \beta_{j1} x_{ij1} + \sum_{k=2}^{4} \sum_{j=2}^{3} \beta_{jk} x_{ijk} + \sum_{j=2}^{4} \beta_{j5} x_{ij5}, \quad i \in S, \quad S \in \Psi_h, \tag{36}$$

where the x_{ijk} are dummy variables defined in the first column of Table 2, and the β_{jk} are the associated regression parameters. Note that we use $j = 1$ as the reference level for each attribute and that $\beta_{j1},\ 1 \le j \le 5$, are brand-specific constants (intercepts) that capture the values of brands B, C, D, and E, respectively, relative to the reference brand A.

Recall that a person has to select at most one (i.e., one or none) of the alternatives from choice set S. Let $P_{0h} = 1 - \phi$ denote the probability that a person makes no choice. The unconditional probability that person h selects alternative $i \in S$ is then given by

$$P_{ih} = \phi \times \frac{\exp(V_{ih})}{\sum_{\ell \in S} \exp(V_{\ell h})}, \quad i, \ell \in S, \quad S \in \Psi_h. \tag{37}$$

Let y_{ih} indicate whether (1) or not (0) alternative $i \in S$ is selected by person h. Then the likelihood function is

$$L = \prod_{h=1}^{N} \prod_{S \in \Psi_h} \left([P_{0h}]^{\left(1 - \sum_{i \in S} y_{ih}\right)} \prod_{i \in S} [P_{ih}]^{y_{ih}} \right), \tag{38}$$

where $\sum_{i \in S} y_{ih} = 0$ if no item is chosen from S; and $\sum_{i \in S} y_{ih} = 1$, otherwise.

The maximization of the likelihood function in equation (38) produces the following utility (unobserved response) function:

$$\hat{V}_{ih} = -0.35x_{i21} - 0.58x_{i31} - 0.836x_{i41} - 1.181x_{i51} + 1.48x_{i22} + 2.16x_{i32} + 0.456x_{i23} + 0.82x_{i33} + 0.66x_{i24} + 0.91x_{i34} - 0.21x_{i25} - 0.85x_{i35} - 1.22x_{i45}, \; LL = -2617.75,$$

with $\hat{\phi} = 0.88$. These estimates, which are all significant at the $p = 0.05$ level, do not appear to suggest a subset-conjunctive rule.

Table 2 shows the results for each of the five subset-conjunctive models. The goodness-of-fit, assessed in terms of the likelihood value, is the best for a subset conjunctive rule with $t = 4$, which

[10] As this is a proprietary study, we were not provided with the actual brand names and the actual price levels used in the study.

TABLE 2.
Subset-conjunction consideration probabilities for personal computers.

Variable	Subset size (t) Conjunctive 5	4	3	2	Disjunctive 1
Brand					
x_{11} A	1.00	0.74	0.57	0.50	0.25
x_{21} B	0.99	0.55	0.34	0.30	0.10
x_{31} C	0.95	0.46	0.25	0.20	0.04
x_{41} D	0.83	0.30	0.08	0.07	0.00
x_{51} E	0.72	0.23	0.00	0.00	0.00
Performance					
x_{12} Below average	0.34	0.02	0.00	0.00	0.00
x_{22} Average	0.80	0.64	0.61	0.59	0.28
x_{32} Above average	1.00	1.00	1.00	1.00	0.53
Warranty					
x_{13} 90 day	1.00	0.58	0.12	0.03	0.00
x_{23} 1 year	0.77	0.84	0.31	0.19	0.00
x_{33} 5 year	1.00	1.00	0.51	0.36	0.11
Service					
x_{14} Ship back to mfg	0.70	0.52	0.31	0.03	0.00
x_{24} Service at local dealer	0.97	0.91	0.67	0.30	0.08
x_{34} On-site service	1.00	1.00	0.78	0.43	0.17
Price					
x_{15} Low	1.00	0.97	1.00	0.52	0.25
x_{25} Med–Low	0.98	0.87	0.89	0.42	0.20
x_{35} Med–High	0.79	0.55	0.53	0.11	0.01
x_{45} High	0.63	0.39	0.33	0.00	0.00
Minus log-likelihood	2641.10	2600.32	2619.50	2644.23	2801.43

Subset size $t = 4$ is best solution.

has a log-likelihood value slightly lower than the value obtained for the nested logit choice model. The disjunctive model has by far the worst fit and stands apart from the others. The conjunctive and subset-two models have comparable values for the log-likelihood. Both are substantially worse than the subset-three and subset-four models. As one might expect, the consideration probabilities for the subset conjunctive rule with $t = 3$ are smaller than the corresponding values for $t = 4$, because the former requires an acceptable alternative to "qualify" on fewer attributes.

For $t = 4$, there is substantial separation in the consideration probabilities across the levels of each attribute, implying that all attributes matter for consideration. The attribute with the largest range over its levels is performance (0.98), followed by price (0.58), brand name (0.51), service (0.48), and warranty (0.42).

To further assess the goodness-of-fit for the $t = 4$ solution, we compute the values of the mean absolute deviation (MAD) between observed choice proportions and predicted choice probabilities across all choice sets and consumers (see Table 3). By design, each level of the performance, service, and warranty attributes appears in one of the three brands in each choice set. Consequently, the MAD values for these three factors, which range from 0.002 to 0.010, are identical to the overall MAD values. The MAD values for the other two factors are also quite small, ranging between 0.002 and 0.018 for brand names, and between 0 and 0.019 for price. Overall, the $t = 4$ model appears to fit the choice data very well.

TABLE 3.
Mean absolute deviation between observed choice proportions and predicted choice probabilities by subset conjunction model ($t = 4$).

Factor	Number of observations	Mean absolute deviations			
		Alternative 1	Alternative 2	Alternative 3	No-Choice
Brand					
A	1549	0.017[a]	0.003	0.010	0.004
B	1528	0.002	0.004	0.005	0.006
C	1558	0.017	0.006	0.008	0.016
D	1553	0.017	0.002	0.016	0.003
E	1559	0.004	0.002	0.016	0.018
Price level					
Low	1914	0.009[b]	0.000	0.009	0.000
Med–low	1906	0.013	0.007	0.001	0.019
Med–high	1918	0.011	0.000	0.001	0.012
High	1915	0.009	0.000	0.001	0.008
Overall	2608	0.010	0.002	0.002	0.009

Note: The mean absolute deviations for the other factors (performance, service, and warranty) are identical to the overall values reported above, because the design forces each level of these attributes to appear in each choice set.

[a] A person can choose at most one of three alternative personal computers in a choice set. There are 1549 choice sets featuring brand A as one of the three alternatives. Across these choice sets, the difference between the observed choice proportion and predicted choice probability has a mean absolute value of 0.017 for the alternative labeled 1.

[b] There are 1914 choice sets featuring an alternative at a low price. Across these choice sets, the difference between the observed choice proportion and predicted choice probability has a mean absolute value of 0.009 for the alternative labeled 1.

For comparison, we also compute the MAD values for the nested-logit model. The results show that both models produce virtually identical results. Specifically, the overall MAD between the observed choice proportions and predicted choice probabilities ranges between 0 and 0.005 across choice alternatives. The MAD values by factor range between 0.0044 and 0.011 for brand names, and between 0.006 and 0.010 for price.

To illustrate how the results from the subset-conjunctive model offer insights beyond those produced by the nested-logit model, assume that the model with $t = 4$ describes consumer consideration, and only brands A, C, and E are available. Suppose a new product were to be launched with above average performance, five-year warranty, and on-site service. Then this product will be considered by a person only if it is acceptable on one or both of price and brand name. Suppose the product were to be introduced at the lowest price level. Then the choice of a brand name will not be especially important, because brand name has only a small effect on the consideration probability; see Figure 1. Even at the Med–Low price, the consideration probability exceeds 0.90 for all three brand names. But if the product were launched at the high price, then the choice of a brand name becomes an important factor, the consideration probabilities ranging between 0.84 if the product is marketed as brand A and 0.53 if it is marketed as brand E.[11] Thus, the sensitivity of the consideration probability to price changes depends on the choice of brand name. Such inferences are not possible with a nested-logit model, because it does not distinguish

[11] The consideration probability for an alternative is the probability that the alternative is acceptable on at least $t = 4$ attributes. We use equation (17) to compute the consideration probabilities.

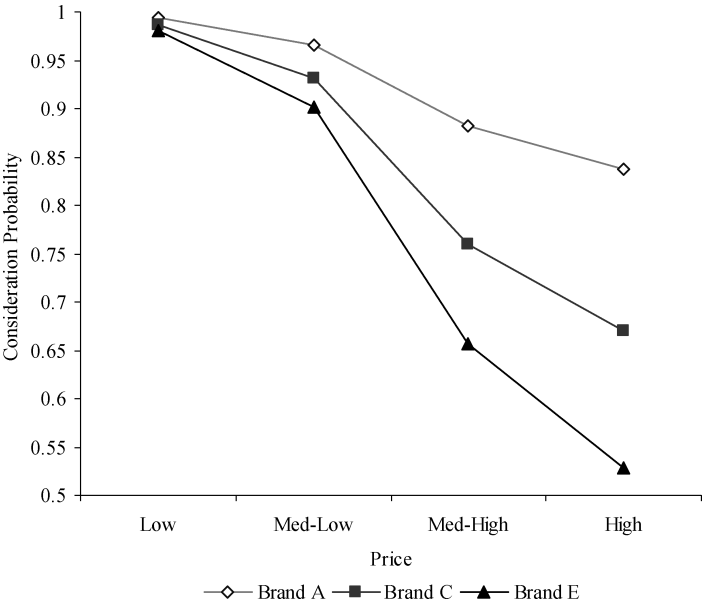

FIGURE 1.
Price impact on consideration probabilities for personal computers.

between product consideration and choice, and because it has additive terms that do not suggest such interaction effects.

The nested-logit and subset-conjunctive models also differ in their market-share predictions. Suppose consumers are offered brands A, C, and E, and each brand has above-average performance, five-year warranty, and on-site service. We examine how the choice probability—and thus the market share—of a brand varies with its price, assuming that the other two brands are available at a Med–Low price. Figure 2 plots the market shares predicted by the nested-logit and subset-conjunctive models; the two models make very different predictions. Table 4 shows the variation in market shares as the prices of the brands change from high to low. Consider brand A: the nested-logit model predicts it will gain 29% market share, 19% from brand C and 10% from brand E; the subset-conjunctive model predicts it will gain 9% market share, drawing about equally from both brands C and E. Now consider brand C: as its price goes from high to low, its market share increases by 21% according to the nested-logit model and by 17% according to the subset-conjunctive model. According to the subset-conjunctive model, both brands A and E lose about the same market share to brand C; according to the nested-logit model, brand A loses the most. Finally, the nested-logit model predicts a much smaller market-share gain for brand E than does the subset-conjunctive model (13% versus 23%) when the price of brand E is reduced from high to low; the sources of share gain also differ in the two models. Note that these differences in the market-share predictions occur despite the comparable predictive validity of the two models on such measures as likelihood value and MAD. A priori, one cannot say which model is correct, because this requires observing how the market responds to actual price changes. But if these changes were to be observed, one could discriminate between the two models.

4. Simulated Testing of Subset-Conjunctive Models

We performed three simulation experiments to test the proposed class of subset-conjunctive models. Our primary purpose in the first experiment is to assess how well the estimation procedure

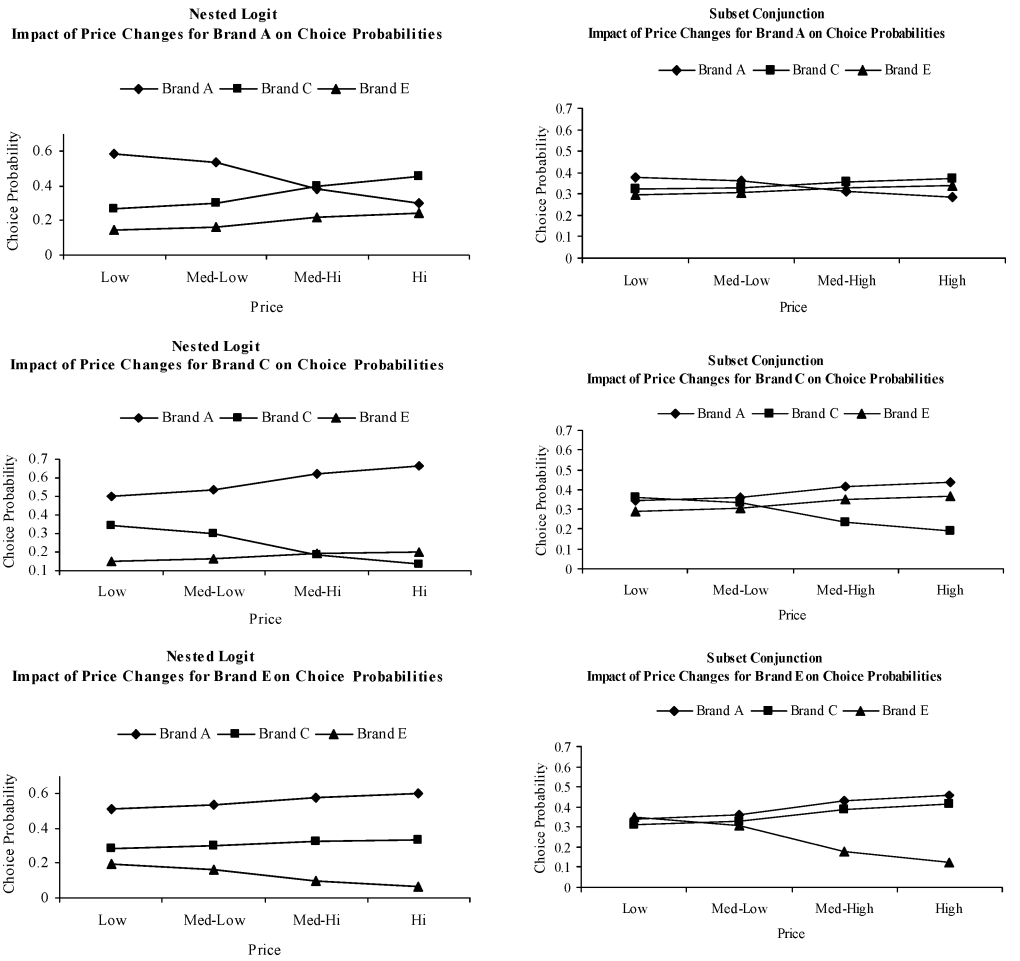

FIGURE 2.
Impact of price changes on choice probabilities.

recovers both the true subset size of the conjunctive process, and the true parameters of the subset-conjunctive models. The second experiment compares the likelihood function values of the correctly specified subset-conjunctive model and of a main-effects logistic regression, for varying subset sizes. Recall that subset-conjunctive models can be represented as additive logistic regression models in the absence of error. Hence the purpose of this experiment is to examine whether correctly specified conjunctive models result in improvement of fit relative to logistic regression models. Finally, in the third experiment, we investigate the incidence of local optima when estimating each model.

4.1. First Simulation

We estimated subset-conjunctive models using a $(2 \times 2 \times 2)$ factorial design with the following treatments: true model (binary vs. multinomial subset-conjunctive), sample size (1600 and 3200 observations), and amount of uncertainty in data (low, high). Recall that an attribute is informative (low uncertainty) if the probabilities of acceptance of its levels are extreme (close to zero or one) and different from each other. One way to operationalize the amount of uncertainty

TABLE 4.
Effect of maximum price change in each brand's price on market shares.

Price change in brand	Predicted change in market share for brand					
	Nested logit			Subset conjunction		
	A	C	E	A	C	E
A	0.29[a]	−0.19	−0.10	0.09	−0.05	−0.04
C	−0.16[b]	0.21	−0.05	−0.10	0.17	−0.08
E	−0.08	−0.05	0.13	−0.12	−0.11	0.23

[a]If the price of brand A changes from "high" to "low," then its market share increases by 29%.
[b]If the price of brand C changes from "high" to "low," then the market share of brand A decreases by 16%.

in a model is by drawing the parameters from Beta distributions with different parameters. We selected Beta(0.1, 0.1) and Beta(0.5, 0.5) for the low and high levels of uncertainty, respectively.

This experiment uses $m = 5$ binary predictors $x_k \in \{0, 1\},\ 1 \leq k \leq 5$, and assumes a subset-conjunctive process with $t = 3$. We generate $N = 1600\,(3200)$ alternatives, i.e., 50 (100) per cell of the 2^5 factorial design, obtained by taking all combinations of the predictor variables. For each alternative $i = 1, \ldots, N$, we used the values of the π_{jk} parameters, generated from the appropriate Beta distributions, to compute the probability π_i^3 that the alternative is acceptable on at least three of the five attributes (see equation (17)). This gives the probability with which we generate the response variable $y_i \in \{0, 1\}$ for the binary subset-conjunctive model.

To generate data for the multinomial subset-conjunctive model, we constructed choice sets of size three. The alternatives in each choice set are randomly drawn (without replacement) from the set of N alternatives. For each choice set, we first compute the considerat probabilities $\pi_1^3, \pi_2^3, \pi_3^3$ and then convert them into choice and no-choice probabilities (see equations (25) and (31)) which we used to generate the choice outcome (y_1, y_2, y_3, y_4), where $y_4 = 1$ indicates no-choice.

In this experiment, we perform 20 replications per treatment. Each replication uses a set of π_{jk} values selected from the appropriate Beta distributions. The data set for each replication is used to estimate subset-conjunctive models with t varying from 1 to 5. The model performance criteria of interest are the recovery of the true model parameters and the recovery of the true subset size, which we set to $t = 3$. We use the MAD between the true and estimated parameters as a measure of bias. We estimate the subset size by estimating the model for $t = 1$ to 5. We select the subset size t, which corresponds to the solution with the maximum likelihood value. Thus the percent of replications that points to $t = 3$ is our measure of true subset size recovery.

The results in Table 5 show that the recovery of the parameters is excellent across all the treatment conditions. In general, the algorithm estimates the parameters accurately for both models, regardless of sample size or amount of uncertainty. The MAD ranges from 0.007 to 0.09, and has a mean value of 0.035. The recovery of the true subset size is also excellent. The percent of replications pointing to the true subset size ($t = 3$) ranges from 90% to 100%, and has a mean of about 97%.

4.2. Second Simulation

The objective of this simulation is to compare the likelihood values of correctly specified subset-conjunctive models with those obtained from logistic regressions. In this experiment, we estimate subset-conjunctive and logistic regression models using data generated using a

TABLE 5.
Summary results for simulation study 1.

	Sample size	Small		Large		
Conjunctive model	Uncertainty	Low	High	Low	High	Overall
	MAD	0.008[a]	0.091	0.007	0.050	0.039
Binary	Selection	100.0%	95.0%	100.0%	95.0%	97.5%
	MAD	0.013	0.055	0.008	0.050	0.031
Multinomial	Selection	100.0%	90.0%	100.0%	95.0%	96.3%
	MAD	0.010	0.073	0.007	0.050	0.035
Overall	Selection	100.0%	92.5%	100.0%	95.0%	96.9%

[a] Across all 20 replications in this cell, the difference between the true and estimated parameters has a mean absolute value, MAD = 0.008. Recall that the subset conjunctive model is estimated while varying the subset size t from 1 to 5. Selection represents the percent of runs whose likelihood values are maximum at $t = 3$. Hence, in this cell, 100% of the runs pointed to the true subset size of $t = 3$.

$(2 \times 2 \times 5)$ factorial design with the following treatments: sample size (1600 and 3200 observations), amount of uncertainty in data (low, high), and subset size (1, 2, 3, 4, and 5). We generate data for the subset-conjunctive model following the same approach discussed above. However, unlike experiment 1 where $t = 3$, we varied the subset size from $t = 1$ to $t = 5$ and generate 100 replications per treatment. Each replication uses a set of π_{jk} values selected from the appropriate Beta distributions. The data sets for each replication are used to estimate both the correctly specified model and a main-effects logistic regression. As our primary goal is to compare likelihood values, we report the relative difference between the log-likelihood values of the two models $(\ln L_{LR} - \ln L_{SC})/\ln L_{SC}$, where the subscript denotes logistic regression (LR) or subset conjunction (SC). Note that this ratio is positive if the subset-conjunctive model has a higher likelihood value than the logistic regression model.

Table 6 reports the results. Across all replications and treatments, the true subset-conjunctive models always produce likelihood values slightly higher (i.e., better) than are obtained from logistic regressions. The relative differences range from as low as 0.001 to as high as 0.0299, and have a mean value of 0.011. These differences tend to be higher with lower amounts of

TABLE 6.
Summary results for simulation study 2.

	Small sample		Large sample		
Subset size	Low uncertainty	High uncertainty	Low uncertainty	High uncertainty	Overall
1	0.019[a]	0.003	0.019	0.002	0.011
2	0.029	0.007	0.024	0.006	0.017
3	0.022	0.003	0.009	0.003	0.009
4	0.025	0.009	0.024	0.007	0.016
5	0.004	0.003	0.004	0.002	0.003
Overall	0.020	0.005	0.016	0.004	0.011

[a] Across all 100 replications in this cell, the relative difference between the log-likelihood value of the subset-conjunctive model and that of a main-effects logistic regression is 0.019.

uncertainty. The average relative difference varies from 0.004 for the high uncertainty level to 0.018 for the low uncertainty level. In other words, with higher uncertainty, it becomes harder to distinguish between correctly specified subset-conjunctive models and logistic regressions. If we assume that empirical data contain more error, the benefits of subset-conjunctive models should be assessed in terms of improved interpretability, and less emphasis should be put on model fit when the likelihood values are close.

4.3. Third Simulation

To investigate the incidence of local optima, we generate 30 data sets (10 for each of the 3 subset-conjunctive models) following the same approach as in experiment 1. However, we specify the average values for the sample size and uncertainty levels. Specifically, we set $N = 2400$ observations and draw the parameters from Beta$(0.3, 0.3)$. For each data set, we estimate the correctly specified model 10 times using random starting values, and once using the true parameter values as starting points. We classify a randomly started solution as a local optimum if its likelihood and parameter estimates do not match those obtained from a solution that used the true parameters as starting values. Using this criterion, none of the 300 runs converge to local optima. Specifically, the mean absolute deviation between the two sets of parameter estimates is zero and the likelihood values are virtually identical. Thus, local optima do not seem to be a problem for correctly specified models. However, as discussed in section 2.2, it is always useful to rerun the estimation procedure several times in any application to ensure that proper convergence is achieved.

In summary, the simulation results suggest that the parameter estimates for the subset-conjunctive models are robust. In general, the accuracy of estimating the parameters and the subset sizes are excellent. Comparisons of model fit show that logistic regressions produces likelihood values very similar to those of a correctly specified subset-conjunctive model, and that the differences in log-likelihood values get smaller with higher uncertainty. Finally, we did not encounter any problems of convergence, for the sample size and the uncertainty level examined.

5. Conclusion

One often encounters "logical" varieties of noncompensatory rules in decision making. Doctors use such rules to diagnose illnesses, marketing executives use them to describe target markets, and consumers use them to screen alternatives into consideration or choice sets. We introduce a generalization of disjunctive and conjunctive decision rules in which an acceptable alternative satisfies a minimum number of criteria, not necessarily one criterion or all criteria. The data produced by a subset-conjunctive rule result in a confounding of main and interaction effects in a logistic regression, provided the error in the responses is small. With greater response error, a logistic regression gives parameter estimates that do not reflect the underlying decision process, even though they produce good fits to the data. Thus, one cannot use a logistic regression to infer a subset-conjunctive rule from classification data. To make such inferences, or to test the consistency of data with a subset-conjunctive rule, we propose a probabilistic form of a subset-conjunctive strategy in which an attribute level has a probability of being acceptable to a person. We describe an extension of the model and propose a choice model in which consideration is modeled in the first stage as a subset-conjunctive strategy, and choice is modeled in the second (conditional) stage as a function of the consideration probabilities. We describe methods for estimating the unobserved probabilities using binary and multinomial choice data, and illustrate the models using cancer diagnosis and consumer psychology examples.

Computationally, standard nonlinear optimization packages, such as those available in SAS, can be used for estimating model parameters for 20–30 dummy coded predictor variables. In all

our examples, these procedures converged in a few seconds, and never took more than 5 minutes, on a mainframe computer. For larger problems, it is preferable to use specialized algorithms, several of which are described in the Appendix. All of these are approximation algorithms, which can be used to obtain good starting solutions for an exact solution procedure.

We also report the results of simulations that suggest the adequacy of the estimation procedure in recovering both the true parameter values and the subset sizes. Local optima do not appear to be a problem for correctly specified models. We suggest using multiple starting values to further guard against local optima. The simulation results show that logistic regression models have likelihood values close to the likelihood values of correctly specified subset-conjunctive models and that the differences between the two models get smaller as the amount of uncertainty in data increases. This result emphasizes the need for assessing the benefits of subset-conjunctive models in terms of improved interpretability, and not on model fit alone.

One limitation of the proposed models is their restriction to discrete attributes. A useful extension is to allow a mixture of continuous and discrete attributes; for all continuous attributes, Einhorn's (1970) model is a reasonable approximation for conjunctive/disjunctive strategies. A second limitation is the inability of the present model to reflect differences in logical strategies used by different groups of individuals. While this is not a significant consideration in the breast-cancer example, it is an important concern in the consumer psychology example. Latent-class or Bayesian extensions of our methods are can be useful for reflecting heterogeneity in processes and probabilities.

6. Appendix. Parameter Estimation for Binary Response

We restrict the discussion to the estimation of a conjunctive model ($t = m$). The estimation of a disjunctive model follows the same approach; one only has to reverse the code of the response variable.

We begin by substituting for π_i^m and taking the logarithm of the likelihood function

$$L_m = \prod_{h=1}^{N} \prod_{i=1}^{N_h} \left(\pi_i^m\right)^{y_{ih}} \times \left(1 - \pi_i^m\right)^{1-y_{ih}}.$$

This gives

$$\log L_m = \sum_{h=1}^{N} \sum_{i=1}^{N_h} y_{ih} \sum_{k=1}^{m} \sum_{j=1}^{n_k} x_{ijk} \log \pi_{jk} + \sum_{h=1}^{N} \sum_{i=1}^{N_h} (1 - y_{ih}) \log(1 - \prod_{k=1}^{m} \prod_{j=1}^{n_k} \pi_{jk}^{x_{ijk}}).$$

The first term on the right-hand side can be simplified by noting that

$$\sum_{h=1}^{N} \sum_{i=1}^{N_h} y_{ih} \sum_{k=1}^{m} \sum_{j=1}^{n_k} x_{ijk} \log \pi_{jk} = \sum_{k=1}^{m} \sum_{j=1}^{n_k} \Big[\sum_{h=1}^{N} \sum_{i=1}^{N_h} y_{ih} x_{ijk} \Big] \log \pi_{jk} = \sum_{k=1}^{m} \sum_{j=1}^{n_k} N_{jk}^{A} \log \pi_{jk},$$

where $N_{jk}^{A} = \sum_{h=1}^{N} \sum_{i=1}^{N_h} y_{ih} x_{ijk}$ is the number of acceptable alternatives in the data that have level j of attribute k, $1 \le j \le n_k$, $1 \le k \le m$. However, no such simplification is possible for the second term. One way to construct an approximation algorithm is to replace this second term in $\log L_m$ by a polynomial approximation, after substituting

$$z = -\prod_{k=1}^{m} \prod_{j=1}^{n_k} \pi_{jk}^{x_{ijk}}$$

in the identity (Gradshteyn, Ryzhik, & Jeffrey, 1994, p. 52)

$$\log(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4 + \cdots = \sum_{s=1}^{\infty}(-1)^{s+1}\,\frac{z^s}{s}, \quad -1 < z \leq 1.$$

Truncating the series after (say) the quadratic term and substituting for $\log(1+z)$ in the expression for $\log L_m$ gives

$$\log L_m \approx \sum_{k=1}^{m}\sum_{j=1}^{n_k} N_{jk}^{A} \log \pi_{jk} - \sum_{h=1}^{N}\sum_{i=1}^{N_h}(1-y_{ih})\left[\prod_{k=1}^{m}\prod_{j=1}^{n_k}\pi_{jk}^{x_{ijk}} + \frac{1}{2}\prod_{k=1}^{m}\prod_{j=1}^{n_k}\pi_{jk}^{2x_{ijk}}\right].$$

The first-order conditions yield

$$N_{jk}^{A} = \pi_{jk}\sum_{h=1}^{N}\sum_{i=1}^{N_h}(1-y_{ih}) \sum_{j,k|x_{ijk}=1}\left[\left(\prod_{s=1|s\neq k}^{m}\prod_{u=1|u\neq j}^{n_k}\pi_{su}^{x_{isu}}\right) + \left(\pi_{jk}\prod_{s=1|s\neq k}^{m}\prod_{u=1|u\neq j}^{n_k}\pi_{su}^{x_{isu}}\right)\right],$$

for all $1 \leq j \leq n_k$, $1 \leq k \leq m$. The solution for this system of $(n_1 + n_2 + n_3 + \cdots + n_m)$ simultaneous equations is subject to the constraints $0 \leq \pi_{jk} \leq 1$, and any a priori orderings of the π_{jk}; such a solution can be found using standard gradient search methods.

Parameter estimates can also be obtained from a scoring model, developed along the lines described by Rao (1973, p. 366) for the multinomial distribution. The score at π_{jk} is

$$S_{jk} = \frac{d\log L_m}{d\pi_{jk}}, \quad \mathcal{I}(\pi_{jk}) = V\left(\frac{d\log L_m}{d\pi_{jk}}\right),$$

where $\log L_m$ denotes the log-likelihood function and V is the variance operator. If the values of the efficient scores and information at the trial values $\pi = \{\pi_{jk}^0\}$ are indicated with the index 0, then small additive corrections $\delta\pi_{jk}$ are given by the simultaneous equations $\mathcal{I}\pi = \mathbf{S^0}$, where $\mathcal{I}$ is the information matrix, π is a column vector of the π_{jk}, and $\mathbf{S^0}$ is a column vector of the S_{jk}. This operation is repeated with corrected values each time until stable values of the π_{jk} are obtained. The variance of the final estimates $\hat{\pi}_{jk}$ is given by the ith diagonal element of $\mathcal{I}$. The major computational step in this method is the inversion of the information matrix at each stage of the approximation. Rao observes that the inverse typically does not change much after the first few iterations, and can be kept fixed after this point.

Finally, methods of moments can be used to estimate the parameters when there are multiple observations per treatment/alternative, and when there are more treatments than the number of parameters. Equating the observed and expected proportions of acceptable alternatives for each treatment gives a system of simultaneous equations, from which the parameters can be estimated using weighted least squares, the weights reflecting the relative sample sizes on which the treatment proportions are estimated.

References

Andrews, R.L., & Srinivasan, T.C. (1995). Studying consideration effects in empirical choice models using scanner panel data. *Journal of Marketing Research, 32*, 30–41.

Barthelemy, J.-P., & Mullet, E. (1987). A polynomial model for expert categorical data. In E.E. Roskam, & R. Suck (Eds.), *Progress in Mathematical Psychology,* Vol. I, Amsterdam: Elsevier Science.

Barthelemy, J.-P., & Mullet, E. (1996). Information processing in similarity judgements. *British Journal of Mathematical and Statistical Psychology, 49*, 225–240.

Ben-Akiva, M., & Lerman, S.R. (1993). *Discrete choice analysis*. Cambridge, MA: MIT Press.

Boros, E., Hammer, P.L., & Hooker, J.N. (1994). Predicting cause–effect relationships from incomplete discrete observations. *SIAM Journal on Discrete Mathematics, 7,* 531–543.

Boros, E., Hammer, P., & Hooker, J.N. (1995). Boolean regression. *Annals of Operations Research, 58*, 201–226.

Coombs, C.H. (1951). Mathematical models in psychological scaling. *Journal of the American Statistical Association, 46* (256), 480–489.

Cooper, L.G. (1993). Market-share models. In J. Eliashberg, & G. L. Lillien (Eds.), *Handbooks of Operations Research and Management Science,* Vol. 5, *Marketing*, (pp. 257–313). Amsterdam: Elsevier Science.

Crama, Y., Hammer, P.L., & Ibaraki, T. (1988). Cause–effect relationships and partially defined Boolean functions. *Annals of Operations Research, 16*, 299–325.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Dawes, R.M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Einhorn, H.J. (1970). The use of nonlinear compensatory models in decision making. *Psychological Bulletin, 73*, 221–230.

Gradshteyn, I.S., Ryzhik, I.M., & Jeffrey, A. (1994). *Tables of integrals, series and products* (5th ed.), San Diego, CA: Academic Press.

Grether, D., & Wilde, L. (1984). An analysis of conjunctive choice: Theory and experiments. *Journal of Consumer Research, 10* (4), 373–386.

Huber, J., & Klein, N. (1991). Adapting cutoffs to the choice environment: The effects of attribute correlation and reliability. *Journal of Consumer Research, 18*, 346–357.

Leenen, I., & Van Mechelen, I. (1998). A branch-and-bound algorithm for Boolean regression. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Data highways and information flooding: A challenge for classification and data analysis* (pp. 164–171. Berlin: Springer-Verlag.

Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Lussier, D.A., & Olshavsky, R.W. (1979). Task complexity and contingent processing in brand choice. *Journal of Consumer Research, 6* (2), 154–165.

Maddala, G.S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*. New York: Academic Press.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64* (2), 187–212.

Mela, C., & Lehmann, D.R. (1995). Using fuzzy set theoretic techniques to identify preference rules from interactions in the linear model: An empirical study. *Fuzzy Sets and Systems, 71*, 165–181.

Montgomery, H. (1983). Decision rules and the search for a dominance structure: Toward a process model of decision-making. In P.C. Humphrey, O. Svenson, & A. Vari (Eds.), *Analyzing and aiding decision process.* Amsterdam: North-Holland.

Payne, J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance, 16*, 366–387.

Payne, J.W., Bettman, J.R., & Johnson, E.L. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534–552.

Rao, C.R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Roberts, J., & Lattin, J. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research, 28*, 429–440.

Swait, J. (2001). A non-compensatory choice model incorporating attribute cut-offs. *Transportation Research Part B, 35*, 903–925.

Teigen, K.H., Martinussen, M., & Lund, T. (1996). Linda versus World Cup: Conjunctive probabilities in three-event fictional and real-life predictions. *Journal of Behavioral Decision Making, 9*, 77–93.

Van Mechelen, I. (1988). Prediction of a dichotomous criterion variable by means of a logical combination of dichotomous predictors. *Mathematiques, informatiques et sciences humaines, 102*, 47–54.

Westenberg, M.R.M., & Koele, P. (1994). Multi-attribute evaluation processes: Methodological and conceptual issues. *Acta Psychologica, 87*, 65–84.

Wolberg, W.H., Street, W.N., Heisey, D.M., & Mangasarian, O.L. (1995). Computerized breast cancer diagnosis and prognosis from fine needle aspirates. *Archives of Surgery, 130*, 511–516.

Wright, P.L. (1975). Consumer choice strategies: Simplifying versus optimizing. *Journal of Marketing Research, 11*, 60–67.

Wright, P.L., & Barbour, F. (1977). Phased decision strategies: Sequels to an initial screening. In M.K. Starr, & M. Zeleny (Eds.), *TIMS Studies in the Management Sciences,* (Vol. 6, pp. 91–109). Amsterdam: North-Holland.

Manuscript received 8 APR 2000
Final version received 21 JUL 2002