MARKETING SCIENCE
Vol. 13, No. 1, Winter 1994
Printed in U.S.A.

CUSTOMER BASE ANALYSIS: AN INDUSTRIAL PURCHASE PROCESS APPLICATION

DAVID C. SCHMITTLEIN AND ROBERT A. PETERSON
University of Pennsylvania
University of Texas at Austin

Customer base analysis is concerned with using the observed past purchase behavior of customers to understand their current and likely future purchase patterns. More specifically, as developed in Schmittlein et al. (1987), customer base analysis uses data on the frequency, timing, and dollar value of each customer's past purchases to infer

- the number of customers currently active,
- how that number has changed over time,
- which individual customers are most likely still active,
- how much longer each is likely to remain an active customer, and
- how many purchases can be expected from each during any future time period of interest.

In this paper we empirically validate the model proposed by Schmittlein et al. In doing so, we provide one of the few applications of stochastic models to industrial purchase processes and industrial marketing decisions. Besides showing that the model does capture key aspects of the purchase process, we also present a more effective parameter estimation method and some results regarding sampling properties of the parameter estimates. Finally, we extend the model to explicitly incorporate dollar volume of past purchases.

Our results indicate that this kind of customer base analysis can be both effective in predicting purchase patterns and in generating insights into how key customer groups differ. The link of both these benefits to industrial marketing decision making is also discussed.

(Estimation and Other Statistical Techniques; Industrial Marketing; Measurement; Promotion)

1. Introduction

To judge from the published literature, it would be easy to conclude that detailed probabilistic analysis of customer purchase patterns in industrial markets is ineffective and/or irrelevant. This is not to say that no such examples exist (a counterexample is Uncles and Ehrenberg's (1990) study of aviation fuel contracts), but they are virtually nonexistent in the overviews provided by Ehrenberg (1988), Greene (1982), Lilien et al. (1992), and Massy et al. (1970).

The stereotypes of industrial markets suggest several reasons for the absence of such analyses. (For a summary of stylized distinctions between industrial and consumer markets see Webster and Wind (1977) or Jackson and Cooper (1988).) One is the lack, in some industrial markets (e.g., nuclear power plant construction), of any appreciable degree of repeat buying over the relevant time horizon. Such a lack of repurchase behavior does not, however, characterize all industrial markets. Indeed, the notion of "relationship

0732-2399/94/1301/0041$01.25

marketing" (Kotler 1991, pp. 8, 205) is based entirely on the importance of repeat buying.

Another plausible reason is the generally smaller number of customers in industrial markets compared to consumer markets. Formal models may provide little value if the relative handful of key customers can simply be telephoned. But this again does not characterize all industrial marketing. Having (say) 20,000 business customers of a potential market of 80,000 firms is certainly smaller than 20 million consumer households out of 80 million in total. But a summary of past and projected customer activity cannot simply stem from speaking personally to each customer, and the salesforce may lack the willingness or ability to undertake such individual analyses.

A third possible explanation is the supposed paucity of data on industrial product customer purchases (relative to supermarket packaged goods). Like the two preceding explanations, this one also contains a grain of truth, but does not apply universally. Indeed, many industrial marketers have a wealth of information detailing past purchases by their customer base (even more detailed information about individual customers than retail marketers). What these firms lack, relative to consumer markets, is (sometimes) systematic information about what *other* firms were offering at the same time, and (usually) systematic information about customer purchases of those other firms' goods/services. Thus they do not have the complete purchase record (and "store environment" information) for a set of households of the type available in consumer nondurables from market research firms such as A. C. Nielsen or Information Resources, Inc.

This data limitation does inhibit some probabilistic modeling applications. In fact, it rules out the logit-type analyses of customer product choice that have become so popular for consumer packaged goods (e.g., Gaudagni and Little 1983; Lattin and McAlister 1985; Zufryden 1986). On the other hand, industrial marketers have an advantage in that their customer base typically represents a *census* of their customers, rather than the *sample* employed by consumer goods firms. This is a major benefit in using individual-level purchase models for direct marketing/targeting decisions. In this respect they resemble a variety of consumer goods businesses having such purchase information for their entire customer base, e.g., many service businesses (medical offices, brokerage firms) and sellers of nonsupermarket goods (department stores for charge card purchases, magazine publishers). Besides this similarity in customer information available, these firms share another important quality, namely the ability to benefit from answers to questions like:

(1) How many active customers does the firm now have, and has the number been growing or declining?

(2) Which individual customers are most likely to represent active customers? Inactive customers?

and probably most important

(3) What level of transactions should be expected next period (e.g., year) by those on the customer list, both individually and collectively?

The answers to these questions have implications for very long-range decisions (valuing the business via the future cash stream of its current customers), intermediate-range actions (how much should be spent to attract new customers and/or retain current customers, i.e., what is the long-term value of a new/current customer), and short-range tactics (which customers should be sent direct marketing incentives to attempt to induce immediate purchases). As indicated previously in Schmittlein et al. (1987, hereafter SMC), these questions are not answered well by aggregate sales data or trends alone.

The model proposed in SMC provides answers to the three questions above for any firms maintaining detailed purchase information (purchase frequency and timing) for their own customer base. For a new industrial customer after initial trial, it envisions a repeat purchase process having the following major features:

(1) Some customers become inactive, i.e., deterministically no longer purchase, at

various points in time. To an observer, the time for which each customer remains active would (if known) appear to have a random component, that is, not all customers become inactive at the same time. Clients may become inactive for any of several reasons, e.g., because they no longer need the product/service being offered, they have switched to another supplier, or they have themselves gone out of business.

(2) While active, customers purchase whenever they wish, with interpurchase times also appearing to have a random component, i.e., purchases do not happen perfectly regularly in time.

(3) When a customer becomes inactive, the selling firm is not notified directly (or at least that information is not recorded in the customer base). Thus, the only (and partial) evidence that a customer may have become inactive is a suspiciously long hiatus since the last observed purchase.

(4) Customers may differ from each other in both their purchase rates (while active) and in their "dropout" rates (i.e., their propensity to become inactive customers). These two rates are treated as varying independently of each other across customers.

The need for a probabilistic modeling approach as in SMC stems most directly from the nonnotification (#3) and random purchasing (#2) features. If the selling firm is generally notified explicitly (as in terminating a long-distance telephone company), then the number of active customers is known directly, without any analysis. Further, even without notification, if purchases were made perfectly regularly while a customer is active, dropout would be evident as soon as this customer's average (and constant) interpurchase time is surpassed. We do not wish to oversell the range of applicability of this paper's modelling effort. Industrial markets are extremely diverse, and certainly many do not possess the characteristics listed above. But on the other hand many industrial markets, including the one to be analyzed in this paper, are well approximated by the four features above.[1]

Although we view the SMC model as an appealing framework for answering the questions posed earlier for many industrial marketers, its original presentation suffered from several drawbacks. The application reported in this paper is intended to overcome what we see as the four most major of these.

1. *Empirical Validation*. No empirical evidence with the SMC customer transaction/retention model has yet been published. We rectify this with our empirical application. The results provide evidence that the model captures the major characteristics of the purchase/dropout process, yields useful insights into the underlying process followed by customers, and has well-calibrated forecasts of future purchase patterns.

2. *Parameter Estimation*. The method suggested in SMC to estimate the model's parameters proved unreliable, so we have developed an improved approach in this paper.

3. *Sampling Properties*. Our bootstrap standard errors allow us to test for the existence of customer subgroups and provide some tentative guidelines for sample size and sample construction (e.g., is it better to use more customers from the available database, or fewer customers for a longer observation period?).

4. *Dollar Volume of Purchases*. The SMC model considers transactions only—not total dollar volume generated. The distinction between the two is particularly severe in industrial markets, where customers do not typically purchase just a single quantity of a standard item offered at a standard price. We extend the SMC model to incorporate dollar volume of purchase.

Besides these specific contributions, we hope to illustrate the potential for application of probabilistic models to the kinds of customer bases that industrial marketers generally actually possess.

[1] Schmittlein et al. (1987) describe in more detail the informational and market/environmental characteristics needed to use this type of model.

Before proceeding to the model development and empirical results we should say a few words about the organization and decision-making context for which this article's application was constructed. The company *manufactures* and *markets* a variety of recordkeeping systems, business forms, stationery, printed forms, and so forth through a nationwide network of independent distributors. These distributors typically have exclusive geographic areas in which to market the firm's products. Moreover, the distributors are contractually prohibited from carrying competing products. For the most part, the firm's product line consists of consumable items, many of which are replaceable with competitors' offerings.

There were several "needs" that collectively motivated this research. Although the firm had complete purchase records on its customers, and although it would archive customers after a certain period of inactivity, it wanted an independent, objective method for estimating the number of customers it actually had. This desire related to two issues the company was attempting to resolve. One issue was whether it should attempt to establish a direct distribution channel to its current customers. In other words, the firm was considering a couple of direct marketing approaches (e.g., mailing catalogs to customers and/or taking their orders directly and/or instituting a telemarketing operation) that would operate independently of (and, in one sense, compete with) the distribution system currently being used. For a direct marketing channel to be effective, it would be important to establish the timing of contacts (i.e., if and when should the customer be sent a catalog), estimate the value of a customer, etc.

The second issue related to the value of specific customer subsets. In particular, distributors effectively "own" their customers or accounts in their geographical area. Sometimes distributors attempted to sell their accounts to other entities or back to the firm or the firm attempted to buy out distributors. Both of these actions require an objective assessment of the value of a particular customer base. At the time of our research, the firm used a rule-of-thumb in arriving at the value of a customer base. This rule-of-thumb typically produced results that were not acceptable to either the distributor or the company and, consequently, all too frequently resulted in litigation.

The modelling effort reported here was designed to address both of these key issues, i.e., coordinating a direct marketing effort, and assessing the value of any particular group of current customers. As background to this application, the next section provides an overview of the customer purchase/retention model proposed in SMC, followed by development of a new parameter estimation method. Subsequently, we present an extension of the SMC model to include explicit analysis of the dollar purchase volume. This applies to industrial applications where a single purchase (order) can involve multiple different products at a nonstandard/negotiated overall purchase price for the order, and serves as the basis for estimating the value of a specific customer or customer group. After describing the industrial customer database available for this study, we present the empirical application for our extension of SMC's transactions model. This includes individual customer purchase forecasts, aggregate forecasts, customer base segmentation, and sampling issues. Finally, we validate the extended SMC model in several ways. First we examine customer reorder patterns to assess the dropout phenomenon assumption above. This assumption regarding the ability to infer active/inactive customer status is then put to a second test via a telephone survey to a subset of the customer base analyzed. Lastly, we cross-validate the model's individual customer-by-customer predictions of the number of reorders, and the total dollar volume of reorders, in a future time period.

2. A Model of Customer Purchase and Retention

The Modeling Perspective

To answer the managerial questions posed in the Introduction, we rely on the observed history of customer transactions with the firm of interest. This history is interpreted

through the prism of the customer transactions/customer retention model described here to provide the desired insights and predictions. Schmittlein et al. (1987) derived the properties we will need for the basic model (ignoring, for the moment, the dollar volume of transactions). They also discuss at length the sort of information and market conditions needed for its application. The two key conditions for this model to hold are that (1) customers can reorder whenever they wish, and (2) the time at which a customer becomes inactive is not directly observable by the supplying firm. Here we will simply review the model's assumptions and list the key equations.

The decision to base customer inferences and predictions specifically on past transaction patterns is a deliberate one. The major reasons are

(1) past purchase patterns are *widely available*,

(2) past purchase patterns tend to be *effective predictors*, and

(3) past purchase patterns can be *rich predictors*.

Point (1) was discussed earlier. The relative ability of past purchase behavior to capture subsequent purchase propensities is well known, especially in direct marketing, where it generally outpredicts other geodemographic customer information available. Similarly, choice modelers often find purchase history to be much more predictive than marketing mix variables such as price or promotions (Fader and Lattin 1993; Guadagni and Little 1983). In spite of this, one might well ask "why not include past purchase, geodemographics, and marketing mix efforts as *joint* predictors of future purchases?" Indeed, we are as interested as anyone in the latter two effects; and, lacking any relevant empirically-validated theory, we would join those adding the latter effects to the right side of ad hoc regression or logit-type predictive models. But we do *not* lack such a theory for the implication of past purchases. Accordingly, we use the theory to detail the unique role played by past purchase patterns both to indicate very specifically their link to inferences and predictions of interest, and to form a rich framework in which various kinds of geodemographic and marketing mix effects can be considered. This brings us to point (3).

It would certainly be possible to calibrate a model predicting each of the following quantities using a sensible set of purchase history information (e.g., purchase frequency, timing of most recent purchase) and some convenient (e.g., logistic) model formulation:

(a) the number of transactions made by a customer in the next T_1 months;

(b) the probability that a customer will make at least K transactions in the next T_2 months, and

(c) the number of months that will elapse until the customer's next transaction.

Unfortunately, calibrating a model for (a) tells us nearly nothing about (b) and (c), and vice-versa. Further, the analysis must be completely redone every time the cutoff K or time period (T_1, T_2) of interest changes.

In contrast, the modeling approach exemplified in SMC provides predictions for all the quantities (a)–(c), for each value (K, T_1, T_2) from a *single* parsimonious set of parameters. In addition, those parameters provide convenient "entry points" for examining the impact of geodemographic or marketing mix variables. Specifically, we can envision these factors affecting a customer's purchase rate, dropout rate, or both. This type of extension to the basic model is in the spirit of Gupta (1988, 1991), Jones and Zufryden (1980), and Wagner and Taudes (1986), all of which incorporate exogenous variables in a richly predictive model responsive to the special role played by past purchase patterns in repeat purchase processes. We will return to such extensions after describing mathematical particulars of the SMC model, and provide an empirical application in the segmentation results to follow.

SMC Model Assumptions

1. *Transactions by Active Customers*. While active, any customer makes transactions with the firm of interest that are randomly distributed in time, with some customer-

specific long-run transaction rate λ. There is substantial support in consumer markets for the effectiveness of this assumption (Ehrenberg 1988; Morrison and Schmittlein 1988). In industrial markets some indirect support is provided by Uncles and Ehrenberg (1990). While active, then, the number of transactions X, made by a customer in time period of length t is a Poisson random variable:

$$P[X = x|\lambda, t] = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \qquad x = 0, 1, 2, \ldots. \tag{1}$$

2. *Individual Customer Retention/Dropout*. Any customer is viewed as remaining "active" with the firm of interest for some (not directly observed) time period of duration τ. After that time (which may, of course, be very long), the customer no longer purchases from this firm. This customer's dropout phenomenon is assumed to occur randomly in time according to some rate μ. That is, the customer, having been active for t months, has a likelihood μ of dropping out (becoming permanently inactive) in the next short time period, regardless of t. Thus, the time spent as an active customer is an exponential random variable, with density

$$f(t|\mu) = \mu e^{-\mu t} : t > 0. \tag{2}$$

Via assumptions (1) and (2) each customer is characterized by two traits: a purchase rate λ and dropout rate μ.

3. *Heterogeneity in Transaction Rates*. Naturally, some customers have high purchase rates while others buy infrequently. This heterogeneity in transaction rates across customers is assumed to follow a gamma distribution, with p.d.f.

$$f(\lambda|r, \alpha) = \frac{\alpha^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda}; \qquad \lambda > 0, \quad r, \alpha > 0. \tag{3}$$

The mean purchase rate across customers is $E[\lambda] = r/\alpha$ and the variance is r/α^2. The parameter r is an index of the homogeneity in purchase rates across customers.

4. *Heterogeneity in Dropout Rates*. Not all customers drop out at the same time, or with the same rate μ. In the SMC model, dropout rates μ are assumed to vary across customers according to a gamma distribution (unrelated to (3) above), with p.d.f.

$$g(\mu|s, \beta) = \frac{\beta^s}{\Gamma(s)} \mu^{s-1} e^{-\beta\mu}; \qquad \mu > 0, \quad s, \beta > 0. \tag{4}$$

The mean dropout rate is $E[\mu] = s/\beta$ and the variance is s/β^2. The parameter s is an index of the homogeneity in dropout rates cross customers.

5. *Rates λ and μ are Independent*. The purchase rate λ and dropout rate μ are assumed to vary independently across customers.

Key Mathematical Results

Several of SMC's formulas will be helpful in estimating the four model parameters (r, α, s, β) and answering the managerial questions raised in the Introduction. For a randomly chosen customer, the expected number of transactions made in T units of time following an initial purchase is

$$E[X|r, \alpha, s, \beta, T] = \frac{r\beta}{\alpha(s-1)} \left[1 - \left(\frac{\beta}{\beta + T}\right)^{s-1}\right]. \tag{5}$$

The variance, across customers, in the number of such transactions is

$$\text{Var}[X|r, \alpha, s, \beta, T] = E[X|r, \alpha, s, \beta, T] - (E[X|r, \alpha, s, \beta, T])^2 + \frac{2r(r+1)\beta}{\alpha^2(s-1)}\left[\frac{\beta}{s-2} - \frac{\beta}{s-2}\left(\frac{\beta}{\beta+T}\right)^{s-1} - T\left(\frac{\beta}{\beta+T}\right)^{s-1}\right]. \quad (6)$$

Two additional results will suffice to provide the desired inferences and predictions. The first is the probability that a customer with a particular observed transaction history is still active at time T since trial. SMC show that this probability depends on the customer's past purchase history only through the number of purchases X and the time t (since trial) at which the most recent transaction occurred. Thus we denote this desired probability as P[Active$|r, \alpha, s, \beta, X, t, T$]. Its formula is given in Appendix A1. This result for P[Active] enables us to compute the expected number of active customers at any point in (calendar) time, and to rank customers with respect to the likelihood that each is still active. P[Active] has some natural application contexts, e.g., in targeting customers who have become inactive recently.

Other applications focus more on forecasts of future purchases. (Note that, at some time, a customer may have only a small chance of still being active, but *if* active, have a high expected future purchase rate.) In these cases we are interested in the expected number of future purchases X^*, to be made in some future period of length T^*, for a customer with observed purchase history (X, t, T) as above. SMC show that this expectation has a very simple form, namely

$$E[X^*|r, \alpha, s, \beta, X, t, T, T^*] = \text{P[Active}|r, \alpha, s, \beta, X, t, T]E[X|r^*, \alpha^*, s^*, \beta^*, T^*] \quad (7)$$

where $r^* = r + X$, $\alpha^* = \alpha + T$, $s^* = s$, $\beta^* = \beta + T$, and the expectation on the right-hand side of (7) is that given earlier in (5). In other words, after seeing history (X, t, T), we only expect to see future purchases if the customer is still active. *If* the customer has not dropped out, the process acts as if it starts over at time T, but with new (updated) parameter values r^*, α^*, s^*, and β^*.

Equation (7) allows us to predict future transaction levels either separately customer-by-customer or for pooled sets of customers. In the latter case, we can forecast the total number of purchases through any future period T^* of interest, for an entire set of *current* customers of the firm. It reveals, in particular, the number of new customers that must be acquired by the firm (e.g., by its salesforce) to reach a given sales target. This is particularly important in industrial marketing, and is *not* directly available from aggregate sales data alone (due to the dropout phenomenon that is not directly observable). SMC include additional properties of the model discussed here, and further discuss the link to specific management decisions.

We also mentioned in the abstract that the SMC model can be used to predict how much longer a customer will remain active. For one just placing an initial order, this is just the aggregate distribution for dropout time, i.e., the mixture of Equation (2) with Equation (4) as prior. This gamma mixture of exponential random variables is a Pareto distribution, with a discussion and its formula given by Morrison (1978) and Schmittlein and Morrison (1983b). After observing a customer for some time, the expected *remaining* time as an active customer is either zero (if the customer is not now active), an event with probability 1 − P[Active], or is simply an updated mixture of Equation (2) with (4); i.e., the same Pareto distribution as above, but with the updated dropout process parameters (s and β) that were given for Equation (7) above.

One final characteristic of the SMC model merits explicit mention. In some business-to-business markets, customers deal with multiple suppliers and switch among them from order to order. One might ask if such switching (not directly observable by the supplying firm) will invalidate the SMC model. That is, while a customer remains active will reorders from *this* supplier still follow the NBD model (Ehrenberg 1988), when other suppliers are also used? Schmittlein et al. (1985) show that the answer to this question is "yes." That is, under some conditions the NBD will be completely unaffected by this switching; and even if such a condition does not hold precisely, the NBD is very robust to switching among multiple suppliers.

3. Extensions of the SMC Model and Methodology

Here, we describe an improved estimation methodology for the SMC model, and then extend that model to incorporate the dollar volume of purchases.

A Two-step Parameter Estimation Method

In estimating the four model parameters (r, α, s, β), SMC proposed a three-step method of moments procedure. That approach relies on the ratio of values for Equation (5) above for time periods of different length. They did not actually apply the approach to any purchase data, and in our experience these ratios for different (T_1, T_2) values are too unstable to provide reliable parameter estimates.

As an alternative, we suggest the two-step method of moments estimation detailed in Appendix A2. It is much more tractable than maximum likelihood estimation, which would require multiple evaluations of the Gauss hypergeometric function for each reorder in the likelihood function—see (A1) and (A2). The first step will extract as much information as possible out of the first moment (i.e., average observed reorders), since lower-order moments usually have better sampling properties. This produces an estimate for three of the four model parameters. The fourth is estimated by fitting a second moment, i.e., the observed variance in reorders. Bootstrap estimates of these parameter sampling variances are reported in the results to follow.

Extending the SMC Model to Incorporate Dollar Volume

We have discussed the importance in industrial markets of extending SMC's model to incorporate explicitly the dollar volume of orders. Equation (7) provides a prediction of future transactions for each customer, and one simple strategy would be to multiply this expectation times the average dollar volume of past reorders made by this customer. Notice that we now characterize each individual customer by *three* latent traits: the transaction rate (λ), dropout rate (μ), and average dollar volume per reorder (θ).

There are two problems with the simple strategy above. The first is what to do with those accounts that happen not to have reordered yet—what average dollar volume per reorder θ, should be expected of such customers? One solution to this first problem is to substitute the average reorder volume for all customers who have reordered, using this overall average $E[\theta]$, as our estimate $\hat{\theta}$ for each customer with no reorders. But using this solution brings us to the second problem with such a simple strategy. Having used the behavior of *other* customers to estimate θ for some customers lacking any reliable $\hat{\theta}$ using past purchases (i.e., those without reorders), why not use this overall $E[\theta]$ to help estimate θ for *any* customers where $\hat{\theta}$ based on past purchases is unreliable (e.g., those with only one or two reorders)? We will propose a model for reorder quantities that enables us to do just this, and so provide a more stable and accurate estimate of future reorder volume θ for each customer.

For any phenomenon with a random component, such as the volume purchased from occasion to occasion for a particular customer, we know that in anticipating future order

volumes, we should expect a regression effect toward the average across-customer volume $E[\theta]$ (Schmittlein 1989). That is, a customer whose observed historical reorder dollar volumes were relatively low was probably a bit "unlucky" on those particular occasions. The expected future volume will generally be greater than the historical average for this customer but still less than the overall population average $E[\theta]$. The opposite occurs for customers with high observed historical order sizes. Further, this statistical regression-to-the-mean effect is more pronounced when the customer's historical average order size is based on only a small number of observations. We propose a model—that can be used in conjunction with the SMC model—to anticipate this regression effect and provide an effective estimator of θ for each customer. Then individual customer volume predictions can be formed by multiplying this estimator by SMC's Equation (7) for transaction frequency.

For a particular customer observed to have made X reorders to date, let Z_i denote the dollar volume of order i ($i = 1, \ldots, X$). Then of course

$$E[Z_i \mid \theta] = \theta \quad \text{for all } i. \tag{8}$$

Our objective is to compute the reverse conditioned expectation

$$E[\theta \mid Z_1, \ldots, Z_X] \tag{9}$$

which will be the expected dollar volume for future reorders by this customer. As SMC did for the transaction/retention process, we use three sets of assumptions that will enable the updating indicated in (9).

(1) CUSTOMER-LEVEL ASSUMPTION. *We assume that the set of Z_i, $i = 1, \ldots, X$ are i.i.d. normal random variables with mean θ and some variance σ_W^2 which is constant across customers. σ_W^2 is thus the within-customer variance in the amount spent across reorders.*

(2) HETEROGENEITY ASSUMPTION. *The average amount spent per order θ is assumed to vary across customers according to a normal distribution with mean $E[\theta]$ and variance σ_A^2. σ_A^2 is thus the variance in average amount spent across customers.*

(3) INDEPENDENCE FROM THE TRANSACTION/RETENTION PROCESS. *The distribution of average amount spent θ across customers is assumed to be independent of the distribution of the transaction rate λ and the dropout rate μ.*

We will motivate these assumptions shortly, but first we present the resulting formula for the conditional expectation of interest (9). With Assumptions (1) and (2), the confidence that one should place in a single observed past order amount Z_i (relative to relying on the population average $E[\theta]$) is the reliability coefficient

$$\rho_1 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_W^2}. \tag{10}$$

Then if only $X = 1$ previous order was observed, the best estimate for θ as in (9) becomes (Schmittlein 1989)

$$E[\theta \mid Z_1] = \rho_1 Z_1 + (1 - \rho_1) E[\theta]. \tag{11}$$

In general, if $X \geq 1$ historical reorders were observed, the reliability of the historical average

$$\bar{Z} = \frac{1}{X} \sum_{i=1}^{X} Z_i \tag{12}$$

is

$$\rho_X = \frac{\sigma_A^2}{\sigma_A^2 + (\sigma_W^2/X)}, \tag{13}$$

and the expected future volume per reorder (9) is

$$E[\theta | Z_1, \ldots, Z_X] = \left(\frac{X\sigma_A^2}{X\sigma_A^2 + \sigma_W^2}\right)\bar{Z} + \left(\frac{\sigma_w^2}{X\sigma_A^2 + \sigma_w^2}\right)E[\theta]. \tag{14}$$

To compute an expected future dollar volume from a customer with a given purchase history (X, t, T, $\bar{Z}$), we multiply the updated volume/reorder estimate (14) by the expected future number of transactions (7).

In arriving at the key Equation (14), Assumptions (1) and (2) were convenient but not necessary. That formula can also be derived from the standpoint of minimizing squared prediction error in θ without the normality assumptions (Gerber 1979, Chapter 6).[2] Finally, Assumption (3) is a "real" constraint in that we need to isolate the updated θ-value in (14) from the updated transaction/retention process in (7) (and keep the mathematics tractable). Its validity may vary from application to application. In the validation section below, we show that independence of reorder dollar volume from typical interpurchase time is tenable for the current firm's customer base. We next describe that base of customers in more detail.

4. An Industrial Customer Database

The firm whose customer base is analyzed here sells a variety of office products to business customers. We examined transactions made through May 1989. At that time this firm maintained purchase records for several hundred thousand customers with whom it had ever done business. Thus, the firm had access to (and retained) detailed information on each transaction made by each customer. Each record in the database details a sale of some quantity of a particular product to a given customer shipped on a certain date for a particular dollar amount. The key *transaction* of interest is not, however, the individual product record but rather *an order* placed by a customer on a certain date. An order frequently includes multiple different products, and although the product choice and quantity decisions could themselves be the focus of interesting modeling, we consider only the overall order and its dollar volume in this paper. These two quantities (order date and dollar volume per order) are the ones that generalize most easily across industrial markets, coincide with the SMC framework, and provide the basis for forecasting the future customer activities of most interest.

Also accompanying each purchase record is that customer's entry date into the database, i.e., the time at which the customer's initial order was placed. This is important since purchase records are kept by the firm covering only a "window" of slightly over three years (generally March 1986 through May 1989 in our case). For customers placing an order within this period, we know all transactions made within the window, and know the date of initial purchase (even if before March 1986), but do not know about any other orders these customers may have placed prior to 1986. Further, for customers that place no orders at all within the window here, no information at all is available (i.e., not even their entry date into the system). These individuals have already been dropped from the current customer database maintained by the firm, using an ad hoc heuristic (i.e., the three-year rule above). Like the firm, we do not see these omitted customers as providing significant future sales—a conclusion also supported by our empirical results below.

[2] The least-squares criterion for generating (14) does not, however, provide an estimate of the entire conditional distribution for θ, which the normal distribution assumption does supply.

Returning to customers entering before March 1986 and included in the database, we need to reset/update the four gamma distribution parameters for each such customer as of March 1986. We do so to include the relevant information on their purchase history prior to that date. Since no information on orders placed in the pre-March 1986 period is available, the purchase rate parameters (r, α) are the same in March 1986 as if an initial purchase had been made then. But the time since the entry date *does* provide information on the dropout parameters (s, β), since it is known that the customer passed through the period (entry date, March 1986) without becoming inactive (since a transaction was made post-March 1986). So the parameter β is increased by the amount of elapsed time between the entry date and March 1986 for these customers. With this updating, modeling and predictions for these individuals in the post-March 1986 period proceeds normally, i.e., in the same fashion as for those customers (entering post-March 1986) whose entire purchase history is observed. For additional discussion of this "left filtering" of event histories which is common in industrial databases see Schmittlein and Morrison (1983a). Approximately 60 percent of the customers in the database analyzed here entered prior to March 1986, and were thus left-filtered.

For this firm the average order size is approximately $100, with a substantial variation both across customers, and within customers across orders placed. Examining each customer's repurchase activity after one (randomly chosen) transaction, the median inter-order time is about seven months.

Consistent with the various purposes and requirements of the analyses to follow, several different subsets of the overall customer base will be used extensively. These customer groups, and the number of customers contained in each, are listed in Table 1. The entire customer base (A) had over 400,000 customers (over 4.5 million individual purchase records). The authors initially had access to a 1/20 sample (B), i.e., every twentieth customer, from this base. Within the sample (B) 17,825 customers (C) were observed for at least one full month. Also within the customer set (B), 1829 customers were not

TABLE 1

Summary of Customer Groups Used in Analyses

	Customer Group	Description	Number of Customers
Universe	A	Total set of customers	400,000+ (4.5+ million purchase records)
1/20 Sample	B	Customers in sample	21,799
	C	Set B customers observed for at least one full month	17,825
	D	Set B customers not left-filtered and having 30+ months of observed purchase history	1,829
	E	Set D customers observed to have placed at least two reorders	1,056
1/10 Sample	F	Customers in sample	40,000+ (450,000+ purchase records)
	G	Set F customers not left-filtered and having 30+ months of observed purchase history	4,050
	H	Random subset of set F customers tracked for extra calendar time	5,030

left-filtered and had 30+ months of observed purchase history following placement of their initial order (D). Finally, 1,056 of the set (D) customers were observed to have made at least two reorders (E), and were therefore used in estimating the distribution of reorder volumes. Subsequently we were given access to a larger (different) 1/10 sample (F). Within this set were 4,050 customers that were not left-filtered, and that had 30+ months of observed purchase history (G). Sets (F) and (G) were tracked from 1986 through 1988. A random subsample of (F), consisting of 5,030 customers (H), was followed further in time, through May 1989.

5. Parameter Estimates

Parameter Values; SMC Model

The four parameters of the basic transactions model were estimated for a representative sample of $N = 4{,}050$ customers (set (G) in Table 1) who each had at least a 30-month observed purchase history. With the least squares two-step estimation described above, the parameter values are:

all customers (30-month history): transaction process: $r = 0.91$, $\alpha = 3.70$;
dropout process $s = 0.21$, $\beta = 0.29$.

From the transaction process parameters, we see that the mean purchase rate while the customer is active, is $r/\alpha = 0.25$ per month, or about three times per year. We should emphasize that even a simple statistic such as this cannot be calculated directly, since we do not know when the customers who drop out have left. An r-value of 0.91 represents a moderate level of heterogeneity in transaction rates across customers (Schmittlein et al. 1992).

The estimated average dropout rate is $s/\beta = 0.72$. For a customer having that rate, the probability that the account becomes inactive in the next month, given that it has not yet become inactive, is $1 - e^{-0.72} = 0.52$. It is highly unlikely that such an account would "survive" beyond a year. But a second characteristic of the dropout rate distribution suggests that few customers are actually near this mean rate of 0.72. Recall that s is an index of homogeneity in dropout rates: a value of $s = 0.2$ is *very* small. (For example, this gamma distribution has much fatter tails than the exponential distribution.) Thus, some customers have very small dropout rates (and are likely to remain with the firm for a long time) and the rest have quite large rates, becoming inactive within months of initial trial.

These inferences—at least for the long-time customers—seem reasonable. The presence of substantial early dropout can also be seen by even a cursory examination of the database. By way of illustration, we reproduce below the records for the first ten customers analyzed:

Customer No.	Entry Date	Most Recent Reorder	Number of Reorders Observed	P[Active]
1	Pre-3/86	4/89	15	.99
2	Pre-3/86	6/86	2	.18
3	5/86	2/89	1	.99
4	8/86	5/87	2	.38
5	10/86	11/88	6	.95
6	12/86	None	0	.15
7	12/87	1/89	4	.91
8	3/87	None	0	.16
9	5/87	6/87	1	.05
10	Pre-3/86	10/86	2	.91

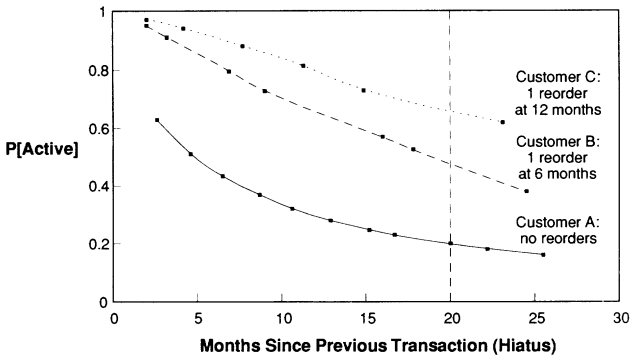

FIGURE 1. Illustrative Relationships Between Order Hiatus and P[Active].

Three customers (#1, 2, 10) were left-filtered. Among this set of customers half seem to be still active (#1, 3, 5, 7, 10) and the other half are probably inactive. This can be seen by examining either the right-most column (the model's P[Active], computed by Equation (A1)) or the customer history itself. All customers appearing inactive based on P[Active] (#2, 4, 6, 8, 9) made 0, 1, or 2 reorders, typically soon after an initial order, and then did nothing for a long time period. In contrast, customers 1, 3, 5, and 7 exhibit a final hiatus that is consistent with their respective average interorder times. Finally, customer 10 is a bit unusual. The account experienced a long hiatus, but is viewed by the model as probably still active. The reason is that this customer has been around for a very long time (not reported in the table above), and thus buying one or more times in the window 3/85–5/89 is sufficient to conclude that the account is still active.

Based on these examples, judging a customer's active/inactive status may seem easy after the fact, i.e., after seeing the value of P[Active] above. But judging it à priori is not nearly as easy; and finding a simple ad hoc index that can be applied automatically in place of the SMC model is even less so. For example, one simple heuristic used by many firms is the duration of the hiatus ($T - t$) since the most recent order. Some cutoff (e.g., a hiatus of two years) is routinely used to characterize a customer as "inactive." (An illustration of this heuristic's use with frequent flyer programs is given in *The New York Times* (1988).)

But such "hiatus rules" are not really effective at capturing major qualitative features of dropout, as we show in Figure 1. Following an initial order, P[Active] does indeed drop with the length of hiatus, i.e., the time since that order for a customer (e.g., customer A in Figure 1) with no reorder. It also drops with hiatus length for customers who repurchased at either six months (customer B) or 12 months (customer C) after an initial order. But *across* customers with *equal* hiatus lengths, P[Active] can still vary greatly. Using our estimated SMC model parameters, when it has been 20 months since the most recent transaction, customer A's probability of still being active is only 0.2, while it is approximately 0.4 and 0.6 for customers B and C, respectively.[3] It does not seem to us an easy matter to construct a simple index likely to replicate the SMC model very well.

In summary, both the parameter values and the illustrative customer histories indicate that while accounts can (and do) become inactive at any time, most customers either drop out relatively early or stay with the firm for an extended period. The four parameters estimates have substantial face validity and are consistent with the sample histories provided above.

[3] Note that the 20-month order hiatus is more "suspicious" in signaling inactivity for customer B than for customer C, since C's transaction rate appears to be lower than B's. In other words, long interorder times are more consistent with customer C's behavior, than they are for customer B. So one would expect P[Active] to be smaller for B than for C.

Sampling Properties and Sample Design Issues

We turn next to some sampling properties of the four SMC model parameters. No evidence on this issue has appeared previously. We first consider the sensitivity of the parameter estimates to the cutoff number of months T_M following initial purchase for which reorders were tracked. The top of half of Table 2 shows the parameter estimates obtained as T_M varies from 12 months to 30 months. We do see some variation in the parameters, mostly reflected in the average dropout rate (s/β) which seems to decline as customers with longer purchase histories are analyzed.

The bottom half of Table 2 shows how the parameter estimates change when T_M is held constant (at 12 months) but the cohort group of customers varies. The order of entry into the customer base proceeds from top to bottom in this part of the table: thus the first row shows estimates using 12 months after trial for the "older" customers—those entering April 1986–November 1986. The bottom row shows estimates for newer customers (that also have 12 months of data following trial). The variation is not substantial, although there may be a slight downward trend in the reorder rate (r/α) among later entrants.

Another important issue is the sampling variation in SMC model estimates. Table 3 provides estimated standard deviations and correlations from 100 bootstrap replicates (Efron and Gong 1983) of the sample in row 1 of Table 2 ($N = 4050$). As a percent of the parameter's value, we note that the average dropout rate (s/β) is the quantity least reliably estimated. (The standard deviation is approximately 47% of the parameter's value.) This makes sense, since our only information about dropout is obtained indirectly, through the time pattern of reorders. We also note, however, that the (high) degree of *heterogeneity* in dropout rates across customers, indicated by the small value of $s = 0.21$ in Table 2, is very reliably estimated. The standard deviation of s across bootstrap samples is only 0.01. The transaction rate parameters while a customer is active (r/α) are also reliably estimated.

Of the 15 parameter correlations listed, only three (r, s), (α, s), and (r/α, α) are not significant at the 0.01 level. As one would expect, the (r, α) and (s, β) correlations are positive, indicating that the estimation is particularly sensitive to attempts to fit the mean transaction rate r/α and mean dropout rate s/β. The other correlation of interest is that between r/α and s/β. The high positive value (0.71) presumably stems from the role that each quantity plays in determining the expected number of transactions made in some time period. That is, a high predicted number of transactions can stem *either* from

TABLE 2

Sensitivity Analysis of SMC Model Estimates: Duration Time Cutoff and Cohort Effects

Months of Data Available	Months of Data Used	Number of Customers*	SMC Model Estimates					
			r	α	s	β	r/α	s/β
30–36	30	4050	0.91	3.70	0.21	0.29	0.25	0.73
24–29	24	3791	1.05	4.73	0.15	0.08	0.22	1.94
18–23	18	3790	0.74	3.78	0.12	0.06	0.20	1.83
12–17	12	3502	1.10	4.28	0.17	0.02	0.26	7.88
30–36	12	4050	1.38	3.51	0.21	0.03	0.39	8.14
24–29	12	3791	2.06	6.02	0.21	0.02	0.34	8.46
18–23	12	3790	1.10	4.17	0.15	0.02	0.26	7.47
12–17	12	3502	1.10	4.28	0.17	0.02	0.26	7.88

* All customers entered on/after March 1986 and so provided a complete record of transactions made in the months following initial purchase. This is Customer Group G mentioned in Table 1.

TABLE 3

SMC Model Sampling Properties: Bootstrap Estimates of Correlations and Standard Deviations

	r	α	s	β	r/α	s/β
r	1.00	0.89	0.06	−0.55	0.55	0.60
α		1.00	−0.09	−0.32	0.12	0.34
s			1.00	0.46	0.30	−0.36
β				1.00	−0.63	−0.95
r/α					1.00	0.71
s/β						1.00
Standard Deviation	0.15	0.48	0.013	0.067	0.017	0.34
Coefficient of Variation	0.16	0.13	0.06	0.23	0.07	0.47

Note: Table based on parameter estimates from N = 4,050 customers tracked over 30 months since placing an initial order. This is Customer Group G mentioned in Table 1.

a high transaction rate while active (r/α) or a low dropout rate (s/β). In fitting the observed number of transactions across bootstrap samples, sometimes the transaction rate r/α is increased, with a counterbalancing increase in the dropout rate s/β.

Finally, we consider one sample issue related to sample design. Calculating SMC model parameters across all transactions of all customers in an industrial customer base can be difficult, time consuming, and unnecessary. Recall that our *subsample* of this firm's customer base (an every 10th customer sample) contained over 40,000 customers and approximately 450,000 purchase records. While we may be interested in *predicting* future potential separately for all 400,000+ customers (not an overly difficult task), we are not sanguine about computing and fitting transaction dynamics across 4.5 million purchase records.

One can think of the sampling frame as having both a "space" (i.e., across customers) and a "time" (amount of time each customer is watched) component. The dataset size—i.e., number of purchase records—goes up by expanding either. Real-world applications

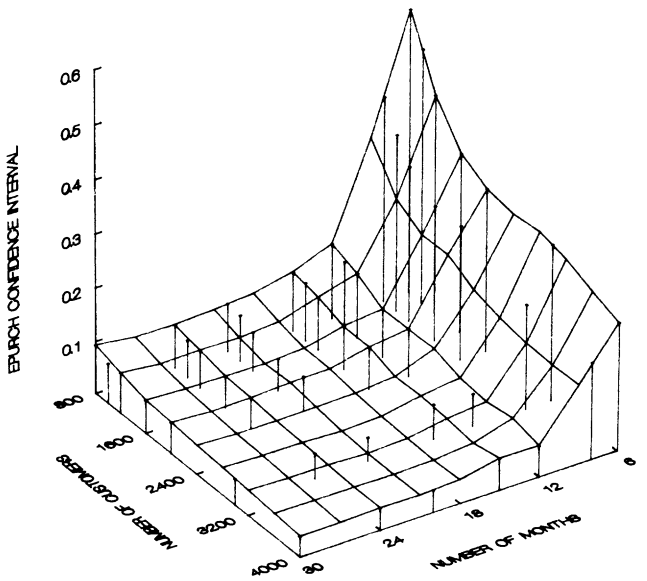

FIGURE 2. Sampling in Space versus Time.

of the SMC model need to make some space/time tradeoff; e.g., would it be better to analyze 12 months of transaction data from 30,000 customers or 30 months of data for 15,000 customers (assuming both produce the same number of purchase records to analyze)?

These kinds of questions can be addressed using the bootstrap method applied earlier. That is, we can see how the standard deviation of a particular parameter of interest changes as the space/time configuration is varied. Of course, this analysis requires actually estimating that parameter in each space/time combination of interest. But doing such an analysis once (or for some hypothetical SMC parameter values) will be sufficient to guide future sample design decisions.

Figure 2 shows the accuracy of (r/α)'s estimate as the # customers/amount-of-time combination varies. Naturally, the standard deviation decreases as each of space and time are expanded. More interestingly, we see that accuracy declines very rapidly (i.e., the confidence interval expands) as the number of months decreases below 12 nearly regardless of the number of customers analyzed. We would conclude here that "good" sample designs for estimating r/α should attempt to include a minimum of 12 months of reorders and 1,600 customers, and then as much additional space/time as computationally feasible.

6. Using the Extended SMC Model

To demonstrate the use of our purchase volume model (14) with SMC's customer transaction/retention model (7), we prepared a variety of five year predictions. The predictions cover the period beginning with the end of our calibration window (May 1989) through April 1994.

The SMC model parameters were those given at the beginning of the last section. In order to extend the model to dollar volume of reorders (e.g., to apply the updating formula (14)), we also need an estimate of σ_W^2, σ_A^2, and $E[\theta]$. The within-customer variance in amount spent across reorders can be estimated separately for each account observed to place at least two reorders. For a sample of 1,056 such accounts (i.e., set (E) in Table 1), the average value of σ_W^2 is 6,059. For these same customers, the variance in θ across accounts is not directly calculable since we do not observe θ. But the total variance in amount spent across both reorders and customers (which equals $\sigma_W^2 + \sigma_A^2$) is available, and equals 11,974. This quantity is computed for the same 1,056 accounts above using one reorder per customer, that reorder being the most recent one in the 30-month period since initial trial. Doing so results in σ_A^2 being estimated as $11,974 - 6,059 = 5,195$. The reliability coefficient ρ_1 is then

$$\rho_1 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_W^2} = \frac{5,915}{11,974} = 0.49. \tag{15}$$

For a customer base where $E[\theta] = \$100$, imagine that a customer was observed to place one reorder at \$150. Using our variance estimates in (14), our best guess regarding this customer's future reorder size is \$125. Similarly, a customer placing one reorder at \$40 has an expected reorder quantity of \$70. Finally, a customer observed to have made *four* reorders with an *average* size of \$150 has an expected reorder size of \$140.

Individual Customer Forecasts

We first show some illustrative customer predictions in Table 4 for the same ten customers we examined earlier. The expected number of reorders is computed using Equation (7). We are able to translate that into a dollar volume projection via (14), listed in the right-most column of the table. Note the dramatic differences across customers in future value due to substantial differences in reorder rates, dropout rates, and reorder sizes. Our expanded SMC model provides a mechanism for combining these various elements of

TABLE 4

5-Year Predictions for Illustrative Customer Accounts

Customer #	Entry Date	Most Recent Reorder	No. of Reorders Observed	*P*[Active]	Expected 5-Year # Reorders	Observed Average Reorder Size ($)	Expected 5-Year $ Volume
1	Pre-3/86	4/89	15	0.99	21.0	70.81	1,521.88
2	Pre-3/86	6/86	2	0.18	0.7	42.50	45.09
3	5/86	2/89	1	0.99	2.8	210.50	424.32
4	8/86	5/87	2	0.38	1.7	59.00	125.48
5	10/86	11/88	6	0.95	10.3	113.00	1,148.68
6	12/86	None	0	0.15	0.3	97.00	24.61
7	12/87	1/89	4	0.91	11.1	54.25	703.49
8	3/87	None	0	0.16	0.3	124.00	29.71
9	5/87	6/87	1	0.05	0.2	79.50	15.52
10	Pre-3/86	10/86	2	0.91	3.9	86.00	350.33

customer history into an overall volume projection, for any time period (and duration) of interest. Note that prediction for another time duration (say, $2\frac{1}{2}$ years) is *not* obtained by merely rescaling the right column of Table 4, due to the progression of customer dropout during the period. One must recompute such a forecast via Equations (7) and (14).

To demonstrate this last point and show the buildup of predicted reorder volume with time, Table 5 lists aggregate predictions during the five year interval discussed above. These expectations are for a set of 5,030 customers (i.e., set (H) in Table 1) whose past reorder patterns formed the basis for those predictions. The internal uses for such aggregate forecasts are obvious. But other less obvious contexts also exist. For example, various legal proceedings sometimes mandate one firm *purchasing* at fair value the current customer base of another firm. For one of several such examples, see *Business Week* (1988). By extension, the model can be applied to any friendly acquisition valuation decision.

Customer Segmentation

There was interest in whether certain à priori-defined customer segments differ in their transaction/dropout processes. In addressing this issue we will illustrate a simple method

TABLE 5

Forecast Volume Pattern Over a 5-Year Period

Number of Years	Incremental Number of Reorders	Cumulative Number of Reorders	Incremental Dollar Volume	Cumulative Dollar Volume	Discounted Cumulative $ Volume*
0.5	3,986	3,986	406,683	406,683	391,006
1	3,883	7,869	395,435	802,118	751,675
1.5	3,805	11,674	387,251	1,189,369	1,086,748
2	3,743	15,417	380,597	1,569,966	1,399,162
2.5	3,688	19,105	374,914	1,944,879	1,691,117
3	3,641	22,745	369,933	2,314,812	1,964,408
3.5	3,598	26,343	365,492	2,680,304	2,220,562
4	3,559	29,901	361,459	3,041,764	2,460,889
4.5	3,523	33,424	357,788	3,399,552	2,686,566
5	3,490	36,913	354,389	3,753,941	2,898,628

* Assumes a 10% discount rate.

for incorporating other explanatory variables (i.e., besides past purchase history) into the SMC model.

The particular à priori segments were defined by type of business conducted, as indicated in the customer's three-digit SIC code. Separate SMC model estimates for three such segments—medical offices, attorneys, and insurance-related films—are shown in Table 6 and allow us to see which, if any, of the four conceptually distinct SMC parameters (i.e., r, r/α, s, and s/β) differ across groups.

An analysis of the Table 6 data was conducted that essentially treats the four SMC parameters as functions of the nominally scaled SIC code in an ANOVA framework. The results reveal that each of the groups is unique in a certain respect. Medical offices have a particularly high transaction rate while active customers ($r/\alpha = 0.664$). Insurance-related firms have moderate transaction rates, but the highest average dropout rates ($s/\beta = 1.62$). Attorneys have the greatest variation across customers in dropout rates: they either quit very soon or stay a very long time. In contrast, attrition among insurance-related firms is more of a gradual and constant occurrence. (Each of these differences across groups is statistically significant at the 0.05 level. Standard errors were estimated for each group via bootstrap samples as in preparing Table 3.)

These differences in dropout dynamics are sketched in Figure 3. Notice that each group rapidly loses a significant fraction of new triers to permanent inactivity. This dropout levels off quickly after 12 months, however. Further the level of dropout varies substantially across the three groups. After two years, 54 percent of the attorneys have become inactive, compared to 64 percent for medical offices and 77 percent for insurance-related firms. Finally, we note that while medical offices become inactive sooner than insurance-related firms, their overall reorder levels are *higher* than those of insurance-related firms through any given time period due to their greater transaction rate (while active).

With respect to explanatory variable effects, we hope this analysis has convinced the reader that:

(1) Purchase history is a special explanatory variable for which a specific model ought to be developed.

(2) Customers can and do differ with respect to *various* aspects of the transaction/retention process. Each of these aspects in turn has its own unique impact on customer predictions of interest, e.g., P[Active] or the expected number of future transactions.

(3) The four SMC model parameters (r, r/α, s, s/β) provide conceptually distinct

TABLE 6

SMC Model Estimates for Different Customer Segments

	Medical Offices	Attorneys	Insurance-related Firms
3-Digit SIC	801—	811—	641—
No. of Customers*	202	307	73
r	1.62	1.61	2.32
α	2.44	6.03	6.65
s	0.211	0.132	0.299
β	0.193	0.114	0.185
r/α	0.664	0.268	0.348
s/β	1.09	1.15	1.62

* To be included in the estimation, a customer must have entered on/after March 1986 and had 24 months of purchase data available.

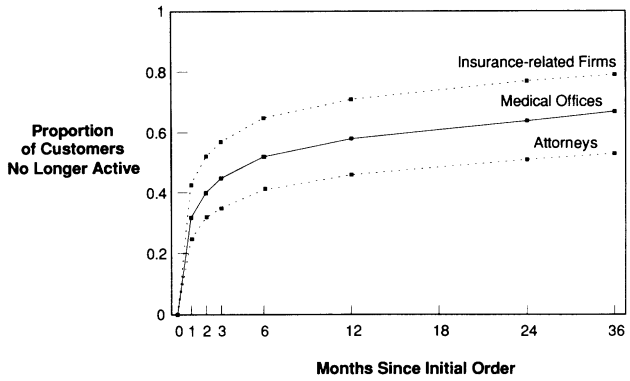

FIGURE 3. Illustrative Customer Dropout Patterns for Three Customer Segments.

and convenient "entry points" (i.e., dependent variables) for the other exogenous variables of interest.

7. Validating the Model Assumptions and Model Predictions

We first take a systematic look at the SMC model's validity. Since the transactions component (i.e., a negative binomial distribution (NBD) model) has a fair amount of empirical support already (see Morrison and Schmittlein (1988) and Uncles and Ehrenberg (1990)), we will concentrate our attention on the customer retention/dropout component. We then will examine our assumption that the simple SMC model's transaction/retention process operates independently from the dollar volume of reorders. Finally, we address the predictive validity of the entire extended SMC model.

Validation of Dropout Phenomenon

Since the dropout phenomenon is not directly observable, its presence (and representation by the SMC model) must be examined through its impact on certain transaction patterns. One such pattern is the trend over (calendar) time in orders placed by a certain set of customers. We consider here a random sample of customers having an initial purchase date prior to April 1988 (i.e., possibly entering before March 1986 as well, and being left-filtered). During the 12 months April 1988–March 1989, the number of orders placed by this static set of customers was:

Quarter	No. of Orders
1. April–June	1,793
2. July–September	1,700
3. October–December	1,670
4. January-March	1,617

Such a drop is suggestive of customers becoming inactive. Of course, this kind of simple statistic may also be influenced by other factors, e.g., a peculiar seasonality in order placements or other calendar-time effects, such as rising prices.

But other analyses tend to rule out these competing explanations and corroborate the dropout hypothesis. First, we note that the special (limiting) case of the SMC model without dropout—i.e., where all customers have a dropout rate $\mu = 0$—is the well-known stationary NBD model (Ehrenberg 1988). We fit the NBD to the observed distribution

for number of transactions in a 30-month period, using data only from customers that *placed* a reorder (i.e., omitting the zero-class). We then use the NBD to predict how large this zero class *should* be, if there is truly no dropout of customers. If the actual number of nonbuyers is greater than the number forecast by the NBD, we have further evidence of the dropout phenomenon.

The actual distribution of reorders, and the NBD model fit, was[4]

	Number of Reorders in 30 Months								
	1	2	3	4	5	6	7	8–10	11+
Observed Number of Customers	253	219	151	136	111	69	72	137	161
NBD Fitted Number of Customers	287	218	171	136	109	89	72	146	182

The NBD predicted number of nonbuyers over this 30 months is 418 customers, which is well below the 520 nonbuying customers observed. Therefore, the (stationary) NBD model substantially underpredicts the number of accounts that do not reorder—again consistent with the dropout hypothesis. Note that the 30-month histogram takes in a longer period of time than the quarterly order pattern presented earlier. Moreover, it covers different calendar time periods for different customers, depending on the customer's entry date.

One remaining alternative explanation for these results would arise from the presence of a substantial subgroup of customers who make an initial purchase and then never consider reordering. In other words, there is a hard-core subgroup that will *never* rebuy, and another (stationary) group that will never drop out (and who follow the NBD model in reordering). This model has some popularity and conceptual appeal (Morrison 1969; Morrison and Schmittlein 1988; Schmittlein et al. 1992), and its extra parameter (relative to the NBD) would indeed enable it to fit the high number of nonbuyers just encountered. Such a Non-User NBD (NUNBD) model does not allow for the continual dropout over time envisioned in SMC. Thus, like the NBD, the NUNBD is a *stationary* model for reorders: the "user" subgroup does not drop out. For any such model the incremental total number of orders placed in each successive time period after initial trial should be approximately equal. This provides a basis for distinguishing the NUNBD from the SMC model, which assumes that incremental orders decrease over time as customers drop out.

The last column of Table 7 reports the incremental order pattern. It shows an unmistakable decline consistent with the SMC model of dropout. Finally, we note that the SMC model does effectively capture these dynamics of the reorder process. Table 7 also shows the model's fit to the buildup of cumulative reorders as time passes since the customer's initial order. For example, by the end of the first year after trial, customers have repurchased 1.65 times on average. By the end of the second year, cumulative repurchases are 2.90—less than double the one-year total. This increasing-at-a-decreasing rate pattern again suggests that dropout is affecting the reorder level. Based on the results reported thus far, we conclude that the SMC model seems to capture effectively the customer transaction/retention process.

Independence of the Transaction Process from the Dollar Value of Reorders

We next validate the extended SMC model's assumption that reorder dollar volume is not seriously affected by reorder timing. Table 8 contains the results of this analysis.

[4] The sample was N = 1,829 customers that entered both on/after March 1986 and before December 1986 (i.e., (D) in Table 1). They could therefore all be observed for 30 months following their initial order. MLE's for the NBD parameters were $r = 0.824$ and $\alpha = 6.00$.

TABLE 7

Cumulative Reorder Levels Following an Initial Order

Time Period Since Initial Order (Months)	Actual Cumulative Number of Reorders Per Customer* Placed Through Period	SMC Model-Estimated Number of Reorders	Actual Incremental Number of Reorders Within Each Three-Month Period
1	0.30	.20	
2	0.43	.37	0.542
3	0.54	.52	
4	0.66	.67	
5	0.78	.80	0.349
6	0.89	.94	
7	1.02	1.06	
8	1.15	1.18	0.374
9	1.26	1.30	
10	1.39	1.42	
11	1.52	1.54	0.383
12	1.65	1.65	
13	1.77	1.76	
14	1.89	1.87	0.351
15	2.00	1.98	
16	2.10	2.08	
17	2.20	2.19	0.318
18	2.32	2.29	
19	2.42	2.39	
20	2.52	2.49	0.290
21	2.61	2.59	
22	2.71	2.69	
23	2.80	2.79	0.291
24	2.90	2.89	
25	2.98	2.98	
26	3.07	3.08	0.263
27	3.16	3.17	
28	3.25	3.27	
29	3.33	3.36	0.258
30	3.42	3.45	

* N = 4,050 customers with initial purchase after March 1986 and before December 1986.

Each of a set of 17,825 customers (i.e., set (C) in Table 1) was monitored following a (randomly chosen) reorder to record the timing of the next reorder (if such an event occurs). Column (b) lists the number of customers whose reorder occurred in month t following the previous order. Column (c) gives the number of customers who had not yet reordered at the start of month t, but who were observed during month t and could have reordered. Thus the *hazard rate* (column (d)) for month t is the fraction of those that could reorder who actually did (i.e., column (b) ÷ column (c)). The C.D.F. in column (e) estimates the proportion of the population that reorders by each month.

TABLE 8

The Reaction of Dollar Volume Per Reorder to Interorder Time

(a) Month	(b) # Customers Reordering	(c) # Customers Who Could Reorder	(d) Hazard Rate	(e) Cumulative Distribution	(f) Avg. $ Amount Per Reorder
1	2,370	17,825	0.133	0.133	112.56
2	1,789	15,866	0.113	0.231	118.34
3	1,257	14,380	0.087	0.298	111.31
4	1,143	13,049	0.088	0.359	116.97
5	915	11,925	0.077	0.409	106.08
6	852	10,871	0.078	0.455	113.09
7	774	9,941	0.078	0.497	109.96
8	713	9,051	0.079	0.537	109.64
9	562	8,318	0.068	0.568	111.11
10	524	7,632	0.069	0.598	120.52
11	500	6,974	0.072	0.627	113.62
12	444	6,377	0.070	0.653	112.97
13	420	5,810	0.072	0.678	110.48
14	298	5,365	0.056	0.696	110.56
15	279	4,938	0.057	0.713	110.49
16	234	4,550	0.051	0.728	116.55
17	171	4,217	0.041	0.739	117.14
18	178	3,909	0.046	0.751	105.72
19	137	3,632	0.038	0.760	110.42
20	121	3,376	0.036	0.769	114.90
21	97	3,141	0.031	0.776	126.31
22	96	2,908	0.033	0.783	97.38
23	63	2,690	0.023	0.788	104.70
24	60	2,470	0.024	0.793	97.78
25	44	2,272	0.019	0.797	95.39
26	38	2,076	0.018	0.801	91.50
27	33	1,893	0.017	0.805	110.42
28	19	1,749	0.011	0.807	127.90
29	20	1,542	0.013	0.809	137.65
30	18	1,323	0.014	0.812	129.83
31	6	1,139	0.005	0.813	96.83
32	6	924	0.006	0.814	71.83
33	16	704	0.023	0.818	105.44
34	10	512	0.020	0.822	91.00
35	2	306	0.007	0.823	117.50
36	2	99	0.020	0.827	140.50

Lastly, column (f) contains the average amount bought per reorder, for orders placed in month t.

Our main observation is that the average reorder amount seems unrelated to the number of months since past purchase: it remains about $110 through the table. This is evidence that at least θ and the transaction rate λ are independent. We also note in passing that the declining hazard rate and the C.D.F. reaching an asymptote below 1.0 are together further evidence of the dropout process as envisioned in SMC.

Predictive Validity: Aggregate Reorder Patterns

The predictive validity of the simple SMC model is illustrated in Figure 4. For the same customers as in Table 7, we estimate the SMC model on the first 12 months of

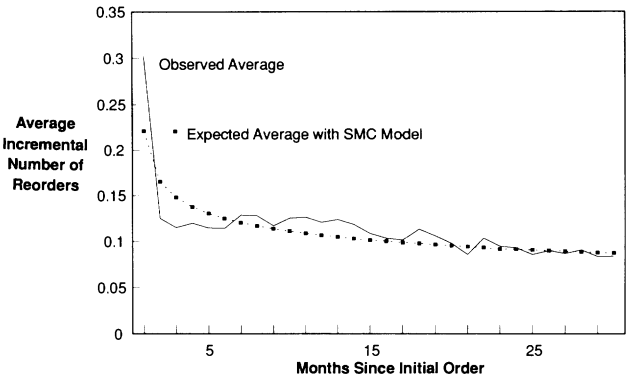

FIGURE 4. Predicting Incremental Orders.

reorder data after initial trial, and then predicted the incremental number of reorders in each of months 12–30. The slow decline in the reorder level shows up in both the model predictions and the empirical results. We next validate the separate customer-by-customer forecasts, in two ways. First, we use a telephone interview to see if customers predicted to still be active (or inactive) by the model in fact see themselves that way. Second, we compare predicted versus actual numbers of reorders, and dollar volume of reorders, for individual customers during a future time period.

Convergence of the Model with Survey Self-report

One of the objectives of the research was to empirically compare the predictions of the SMC model as to the likelihood of a customer being currently active with the customer's self-report of its own activity status. To accomplish this, telephone interviews were attempted with a sample of 400 of the 5,030 customers analyzed earlier. Up to three attempts were made to contact each sampled customer. The interview documented the disposition of the customer with respect to whether it was still active, had gone out-of-business, purchased supplies from a competing company, and so forth. For our purposes, the outcome measure from each interview is essentially a binary indicator: either the customer views itself as still active for this supplier (and hence expects to reorder at some unspecified time in the future) or has become inactive, with no expectation of any future reorders. Our interest is in the association between this self-report measure and the SMC model's probability that the customer is still active. Presumably, a high SMC model probability should be accompanied by a higher propensity for the customer to report an active status.

Of course, neither of these quantities is a perfect indicator of future customer behavior. This is so by definition for the (probabilistic) SMC model. And we know that customer intentions are imperfect predictors of future behavior, i.e., in general not all customers who say they will "definitely buy" will in fact do so (Morwitz and Schmittlein 1992). Thus, relative to future behavior the model and self-report activity status each contain an idiosyncratic error, and will not be perfectly associated with each other. As a result, we would expect the relation between the model's probability and the proportion who self-report an active status to show a positive correlation, but a slope less than one.

This sort of convergent validity is what we find in the results reported in Figure 5. The possible SMC model probabilities were split into deciles, and 40 customers with a model-based probability in each decile were chosen to contact by telephone. The proportion of respondents who report an active status is plotted for each decile. For instance, 12.5% of those customers predicted to have an activity probability between 0.1 and 0.2 actually said they were still active. That percentage self-reporting activity increased to 25% for

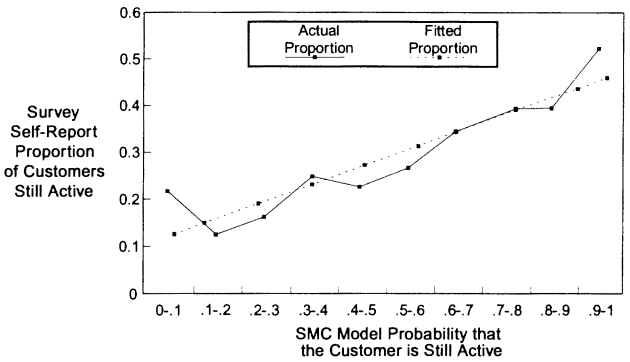

FIGURE 5. Identifying Active Customers: Convergence Between Model Estimates and Survey-based Self-report.

customers with model probabilities between 0.3 and 0.4. Both the raw proportions and the linear regression fit through them show the significant positive correlation and slope-less-than-one hypothesized above. (The regression R^2 was 0.81.) The substantial convergent validity here broadens our confidence both in the SMC model's underlying process assumptions and in its predictions.

Predictive Validity: Individual Customer Transactions

Probably the best single indicator of the model's effectiveness is its ability to predict the frequency and size (dollar volume) of future transactions for individual customers. Values for the seven extended SMC parameters (r, α, s, β, σ_A^2, σ_W^2, $E[\theta]$) were taken to be the estimates reported earlier in the paper. Then, for a different set of customers over a later six-month period, actual reorders were compared to the extended SMC model predictions, those predictions being based on the model parameters and, for each customer, on the purchase history observed to date.[5]

One quantity of interest is the number of reorders a customer placed during the validation period. The mean squared error (MSE) for the SMC model predictions was 0.627. To interpret this number, we also computed predictions for a natural baseline model—namely, a simple extrapolation of that customer's observed history. In other words, the customer's observed reorder rate prior to the six-month validation period was used to predict reorder frequency during the latter period. The MSE for the baseline forecasts was 1.006, indicating a 38% improvement in squared error for the SMC model.

We also validated the actual total dollar volume generated by each customer in the six-month period. Here, the baseline forecast was the customer's observed historical reorder rate (on a six-month basis) multiplied by the average dollar volume per reorder exhibited by the customer to date. The baseline forecast MSE was 19,173 while the MSE for the extended SMC model was 12,884, an improvement of 33% for our model. In short, all of the evidence in this section indicates that the extended SMC model captures the main qualitative characteristics of the customer reorder timing/reorder size/retention process, and also predicts future customer behavior relatively well.

8. Discussion and Conclusions

We have focused on validating and extending the SMC model for customer base analysis. We hope that the family of empirical results presented convinces the reader that the

[5] A subset of customer group H from Table 1 was used, comprised of all who had placed an initial order on/after May 1986 and before November 1988. Predictions were made for the six-month period December 1988–May 1989. A total of 1,606 customers met the criteria, and therefore each had between 0 and 30 months of observed purchase history, to serve as the basis for the six-month prediction.

expanded model is useful for aggregate customer forecasts, customer segmentation, individual customer valuation, and inferences about the reorder/dropout process actually taking place among customers. We have shown how exogenous factors affecting this process can be incorporated in the expanded SMC model. The empirical results provide answers to the key managerial questions raised earlier in the paper.

We think that the application here convincingly indicates that customer dropout can be a real and significant phenomenon and that it can be signaled account-by-account in the pattern of past orders. It is hoped that this paper will help dispel the notion that industrial purchase patterns are not amenable to significant modeling efforts of the general sort that are routine in researching consumer markets. Of course, industrial markets *are* somewhat different and require analyses sensitive to these differences. This is reflected in our current work in two ways: the focus on dollar volume and the use of purchase data only for the firm in question (i.e., ignoring orders placed with competitive suppliers that the firm will not often be aware of in any systematic way).[6]

[6] This paper was received July 16, 1991, and has been with the authors 13 months for 2 revisions. Processed by Scott A. Neslin, Area Editor.

Appendix A1

The Probability that a Given Customer is Still Active

Case 1: $\alpha > \beta$. The probability that a customer is still active, given an observed history of X purchases in time $(0, T)$ since trial, with the most recent purchase at time t, is

$$P[\text{Active}|r, \alpha, s, \beta, X = x, t, T] = \left\{1 + \frac{s}{r+x+s}\left[\left(\frac{\alpha+T}{\alpha+t}\right)^{r+x}\left(\frac{\beta+T}{\alpha+t}\right)^{s} F(a_1, b_1; c_1; z_1(t)) - \left(\frac{\beta+T}{\alpha+T}\right)^{s} F(a_1, b_1; c_1; z_1(T))\right]\right\}^{-1} \quad \text{(A1)}$$

where $a_1 = r + x + s$; $b_1 = s + 1$; $c_1 = r + x + s + 1$; $z_1(y) = (\alpha - \beta)/(\alpha + y)$.

Case 2: $\alpha < \beta$.

$$P[\text{Active}|r, \alpha, s, \beta, X = x, t, T] = \left\{1 + \frac{s}{r+x+s}\left[\left(\frac{\alpha+T}{\beta+t}\right)^{r+x}\left(\frac{\beta+T}{\beta+t}\right)^{s} F(a_2, b_2; c_2; z_2(t) - \left(\frac{\alpha+T}{\beta+T}\right)^{r+x} F(a_2, b_2; c_2; z_2(T))\right]\right\}^{-1} \quad \text{(A2)}$$

where $a_2 = r + x + s$; $b_2 = r + x$; $c_2 = r + x + s + 1$; $z_2(y) = (\beta - \alpha)/(\beta + y)$.

Case 3: $\alpha = \beta$.

$$P[\text{Active}|r, s, \alpha = \beta, X = x, t, T] = \left\{1 + \frac{s}{r+x+s}\left[\left(\frac{\alpha+T}{\alpha+t}\right)^{r+s+x} - 1\right]\right\}^{-1}. \quad \text{(A3)}$$

In (A1) and (A3), $F(a, b; c; z)$ is the Gauss hypergeometric function (Abramowitz and Stegun 1972, p. 588). It can be computed using either numerical integration or the algorithms in Luke (1977). When no transactions are observed in the period $(0, T)$, the probability that the customer is still active is obtained by substituting $X = 0$ and $t = 0$ in (A1), (A2), or (A3).

Appendix A2

A Two-step Estimation Method for the SMC Model

As in SMC, we choose parameter values to best fit the observed number of purchases in a period of duration T predicted by Equation (5). Note that (5) only depends on three of our four parameters $(r/\alpha, s, \beta)$ for a period of any duration T. For $T = 1, 2, \ldots$, then, we have an observed number of purchases per customer and a prediction via (5). We choose $(r/\alpha, s, \beta)$ to minimize the sum (over T) of squared discrepancies in these predictions. Given estimates for these three parameters, we obtain a separate estimate of α (and hence r) by fitting the variance in purchases predicted by Equation (6). As for (5), we minimize (over α, given $r/\alpha, s, \beta$) the sum of squared errors over $T = 1, 2, \ldots$.

Of course, a customer can only provide an observation for the left side of (5) or (6) if the initial purchase was made at least T time periods ago. (That is, a customer "in the system" for only three months cannot provide an estimate of average purchase frequency over a six-month period after trial.) Thus, we will choose some cutoff length of time, T_M, to screen customers for inclusion in the estimation algorithm: only those customers having been in the system for at least time T_M will be used.

The estimation procedure can be summarized as:

Step 1: Choose T_M. For all customers making an initial purchase at least T_M time units ago, choose $(r/\alpha, s, \beta)$ to minimize

$$\sum_{T=1}^{T_M} (E[X \mid r/\alpha, s, \beta, T] - \bar{X}_T(T_M))^2 \tag{A4}$$

where $\bar{X}_T(T_M)$ is the average number of transactions per customer through those customers' first T months. We use an iterated pattern search to minimize (A4).

Step 2: Using Step 1 estimates $\widehat{r/\alpha}$, $\hat{s}$, $\hat{\beta}$; for each $T = 1, \ldots, T_M$, choose $\hat{\alpha}_T$ to equate Var $[X \mid r, \alpha, s, \beta, T]$ with Var $[\bar{X}_T(T_M)]$ where Var $[\bar{X}_T(T_M)]$ is the variance across customers (the same individuals as in Step 1) in the number of transactions per account through those customers' first T months after initial purchase. Average these estimates to obtain $\hat{\alpha}$:

$$\hat{\alpha} = \left(\frac{1}{T_M}\right) \sum_{T=1}^{T_M} \hat{\alpha}_T. \tag{A5}$$

For the main empirical findings of this paper, we use $T_M = 30$ months. The estimation section in the text reports some evidence regarding sensitivity of $(\hat{r}, \hat{\alpha}, \hat{s}, \hat{\beta})$ to choice of T_M.

References

Abramowitz, M. and I. A. Stegun (Eds.) (1972), *Handbook of Mathematical Functions*, New York: Dover Publications, Inc.

Business Week (1988), "Why Pernod Didn't Go Better With Coke," June 20, p. 64.

Efron, B. and G. Gong (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *American Statistician*, 37, 1, 36–48.

Ehrenberg, A. S. C. (1988), *Repeat Buying*, 2nd ed., New York: Oxford University Press.

Fader, P. and J. Lattin (1993), "Accounting for Heterogeneity and Non-Stationarity in a Cross-Sectional Model of Consumer Purchase Behavior," *Marketing Science*, 12, 304–317.

Gerber, H. U. (1979), *An Introduction to Mathematical Risk Theory*, Monograph No. 8, Philadelphia, PA: S. S. Heubner Foundation for Insurance Education, The Wharton School, University of Pennsylvania.

Greene, J. (1982), *Consumer Behavior Models for Non-Statisticians*, New York: Praeger.

Guadagni, P. M. and J. D. C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2, 203–238.

Gupta, S. (1988), "Impact of Sales Promotion on When, What, and How Much to Buy," *Journal of Marketing Research*, 25, 342–356.

——— (1991), "Stochastic Models of Interpurchase Time with Time-Dependent Covariates," *Journal of Marketing Research*, 25, 342–356.

Jackson, R. W. and P. D. Cooper (1988), "Unique Aspects of Marketing Industrial Services," *Industrial Marketing Management*, 17, 111–118.

Jones, J. M. and F. S. Zufryden (1980), "Adding Explanatory Variables to a Consumer Purchase Behavior Model: An Exploratory Study," *Journal of Marketing Research*, 17, 323–334.

Kotler, P. (1991), *Marketing Management: Analysis, Planning, Implementation, and Control*, 7th Ed., Englewood Cliffs, NJ: Prentice-Hall, Inc.

Lattin, J. and L. McAlister (1985), "Using a Variety-Seeking Model to Identify Substitute and Complementary Relationships Among Competing Products," *Journal of Marketing Research*, 22, 330–339.

Lilien, G. L., P. Kotler and S. Moorthy (1992), *Marketing Models*, Englewood Cliffs, NJ: Prentice-Hall.

Luke, Y. L. (1977), *Algorithms for the Computation of Mathematical Functions*, New York: Academic Press.

Massy, W. F., D. B. Montgomery and D. G. Morrison (1970), *Stochastic Models of Buyer Behavior*, Boston, MA: MIT Press.

Morrison, D. G. (1969), "Conditional Trend Analysis: A Model That Allows for Nonusers," *Journal of Marketing Research*, 6, 342–346.

——— (1978), "On Linearly Increasing Mean Residual Lifetimes," *Journal of Applied Probability*, 15, 617–620.

——— and D. C. Schmittlein (1988), "Generalizing the NBD Model for Customer Purchases: What are the Implications and is it Worth the Effort?," *Journal of Business and Economic Statistics*, 6, 145–159.

Morwitz, V. and D. C. Schmittlein (1992), "Using Segmentation to Improve Sales Forecasts Based on Purchase Intent: Which Intenders Will Buy?," *Journal of Marketing Research*, 29, 391–405.

The New York Times (1988), "Mileage Points Can Fly Away," November 20, Travel Section, p. 3.

Schmittlein, D. C. (1989), "Surprising Inferences from Unsurprising Observations: Do Conditional Expectations Really Regress to the Mean?," *The American Statistician*, 43, 176–183.

———, A. C. Bemmaor and D. G. Morrison (1985), "Why Does the NBD Model Work? Robustness in Representing Product Purchases, Brand Purchases, and Imperfectly Recorded Purchases," *Marketing Science*, 4, 255–266.

———, L. G. Cooper and D. G. Morrison (1992), "Truth in Concentration in the Land of (80/20) Laws," *Marketing Science*, 12, 167–183.

——— and D. G. Morrison (1983a), "Modeling and Estimation Using Job Duration Data," *Organizational Behavior and Human Performance*, 32, 1–22.

——— and ——— (1983b), "Prediction of Future Random Events with the Condensed Negative Binomial Distribution," *Journal of the American Statistical Association*, 78, 449–456.

———, ——— and R. Colombo (1987), "Counting Your Customers: Who Are They and What Will They Do Next?," *Management Science*, 33, 1–24.

Uncles, M. D. and A. S. C. Ehrenberg (1990), "Industrial Buying Behavior: Aviation Fuel Contracts," *International Journal of Research in Marketing*, 7, 57–68.

Wagner, U. and A. Taudes (1986), "A Multivariate Polya Model of Brand Choice and Purchase Incidence," *Marketing Science*, 5, 219–244.

Webster, F. E. and Y. Wind (1977), *Organizational Buying Behavior*, Englewood Cliffs, NJ: Prentice-Hall.

Zufryden, F. (1986), "Multibrand Transition Probabilities as a Function of Explanatory Variables: Estimation by a Least-Squares-Based Approach," *Journal of Marketing Research*, 23, 177–183.