Forecasting New Product Sales in a Controlled Test Market Environment

Peter S. Fader
Bruce G. S. Hardie
Robert Stevens
Jim Findley[1]

July 2003

[1]Peter S. Fader is Professor of Marketing at the Wharton School of the University of Pennsylvania (address: 749 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340; phone: (215) 898 1132; email: faderp@wharton.upenn.edu; web: www.petefader.com). Bruce G. S. Hardie is Associate Professor of Marketing, London Business School (email: bhardie@london.edu; web: www.brucehardie.com). Robert Stevens is Vice President Analytic Product Development at Information Resources, Inc. Jim Findley is Senior Vice President Testing Services at Information Resources, Inc.

Abstract

Although new product forecasting is one of the most critical activities for virtually all firms, it tends to be a source of great frustration and indecision for most of them. To address this need, Information Resources, Inc., set out to create a new forecasting system with particular emphasis on new products in consumer packaged goods markets. We document the resulting modeling system, known commercially as IntroCast®, and describe key issues involved in its implementation and managerial interpretations. The model features a "depth-of-repeat" structure, which breaks the new product sales into three underlying components (trial, first repeat, and additional repeat), each of which is modeled independently (including distinct influences for marketing mix effects). We demonstrate the performance of these components — separately and together — for a 52-week forecasting horizon and then validate the model on a number of actual new product tests. The model's excellent performance makes it possible to shorten the timeframe required for model calibration from a typical six-month interval down to a twelve-week period, a very significant improvement for today's highly competitive new product marketplace.

1 Background

Given the cost and risk of launching a new product in the consumer packaged goods (CPG) industry, firms often undertake some form of test market in order to gain a more accurate read on the new product's performance before final decisions about the product launch are made. Historically, this has involved selling the product into retail distribution in one or two markets, observing the product's sales performance for a period ranging from one to two years, and then deciding whether (and how) to further expand distribution coverage in other markets. However, such traditional ("sell-in") test markets have a number of downsides. They are expensive and the costs increase with the duration of the test, as do the opportunity costs of not "going national" earlier. Furthermore, long testing periods give competitors more time to evaluate the performance of the new product and develop effective reactions and preemptive strategies.

Starting with the early work of Fourt and Woodlock (1960), management scientists and marketing researchers began developing a number of models designed to generate forecasts of the new product's longer-term sales performance using panel data collected over the first few months of the test market. Much progress was made in this area for nearly 15 years, but by the late 1970s, academic researchers had turned their attention away to other types of modeling tasks, leaving this area of research largely untouched for the next two decades. At the time these models were being developed, consumer panel data were collected using self-completed paper diaries. Information about merchandising activities, such as coupons and in-store display promotions, were sparse and difficult to measure cleanly. Consequently most models developed in this era did not include the effects of marketing decision variables.

In 1980, the field of test marketing was revolutionized by the introduction of the BehaviorScan® electronic test market service by Information Resources, Inc. (IRI). This service provided marketing managers with unprecedented control over the test market environment. As one of the first applications of UPC scanner data beyond inventory management, the system made it possible to capture point-of-sale purchasing data along with timely, complete, and accurate information about the factors that influence purchasing decisions at the individual consumer level. For the first time, managers could cleanly measure — and control — factors such as retail distribution and merchandising activities, television advertising exposure, and other types of promotions

directed toward consumers. Two key applications of this service are the testing of new products and the testing of advertising; see Lodish et al. (1995) for details of the latter application.

At the time of writing, the BehaviorScan® service is based around five geographically disperse and demographically representative cities. Within each city a panel of several thousand households carry cards that are scanned when checking out from a variety of retail stores, representing over 90% of the grocery dollar volume in those cities.

As marketing managers plan for the introduction of a new product, the BehaviorScan® system gives them tight control over many relevant marketing and environmental variables. IRI maintains its own warehouse and distribution system in these cities, and IRI personnel are in the stores several times each week to assure that the planned distribution, shelf location, and point-of-sale promotional activities are executed as planned. Similar oversight exists for direct-to-consumer promotions, such as samples and newspaper coupons. Finally, addressable cable television technology allows for the targeting of specific advertisements to subsets of the panel households. This high degree of control over the various dimensions of the market environment allows managers to replicate expected conditions in a hypothetical national launch. (Further details of the BehaviorScan® system can be found in Curry (1993) and Larson (1992).)

IRI analysts involved in the generation of new product sales forecasts using data collected via a BehaviorScan® test have employed techniques such as those outlined in Fourt and Woodlock (1960) as well as extensions of that work, e.g., Eskin (1973). But these techniques were designed for use with simple aggregated panel data, not the type of rich individual-level data that is now available. Analysts attempted to extend these original approaches to incorporate the effects of marketing mix variables, but generally did so in a rather ad hoc manner. For instance, certain model parameters from the basic Fourt-Woodlock framework were replaced by polynomial functions of marketing effects. In some cases, these extensions violated "logical consistency" constraints. As but one example, the penetration build, i.e., the cumulative trial rate, for a new product should be a monotonically increasing function bounded between 0 and 100%, but some of the "fixes" to accommodate marketing effects would lead to curves that could, in theory, decrease and/or go outside these bounds.

In the mid-1990s, IRI set out to create a state-of-the-art decision support system to help manage the introduction of a new CPG product. This system, known as IntroCast®, was

designed to be used not only in conjunction with BehaviorScan®-based tests but also with tests conducted in standard panel markets for which there is no controlled distribution. Additionally, the system would be used in the management of regional roll-outs. The managers responsible for this initiative recognized the limitations of IRI's existing forecasting models and decided that the first phase of this decision support development effort should focus on the underlying model of new product sales. This paper reports on the initial results of this development effort, namely a model for forecasting the sales of a new product based on early results from a controlled test market.

2 General Model Structure

Consider the product "Kiwi Bubbles" (a masked name for a real drink product) that underwent a year-long test in two of IRI's BehaviorScan® test markets prior to making the final "go national" decision. Figure 1 plots the purchasing of this product by the 2799 panelists in these two markets for the first 24 weeks of the test.

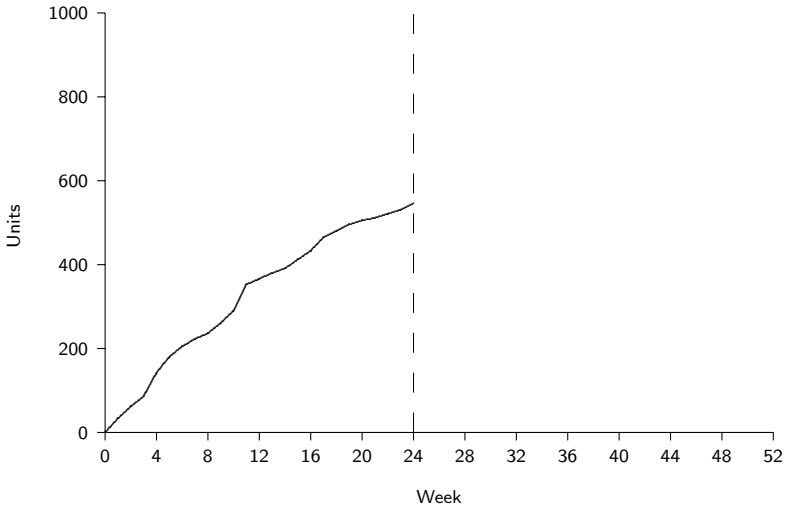

Figure 1: Initial Test Market Sales

Given this information, the brand manager may ask two related questions: (i) "How well is Kiwi Bubbles performing to date?" and (ii) "What sales level can we expect by the end of the year?" Answers to these questions, especially the latter, are key inputs to the final "go-national" decision.

A given overall aggregate sales history could be the realization of very different purchasing scenarios. For example, a low sales level for a new product could be the result of: (i) many consumers making a trial purchase but few of them making a repeat purchase (because the product does not meet their expectations), or (ii) a low trial rate but a high level of repeat purchasing amongst the triers (because the product meets a real need among a relatively small set of buyers). Without a proper trial-repeat sales decomposition, it is impossible to determine which of these scenarios (or the infinite other possible combinations of trial-repeat patterns that could lead to the same aggregate sales curve) best describes the early sales data. Therefore the decomposition of aggregate sales into separate trial and repeat components is central to answering either of these questions. Further insights can be gained by decomposing repeat sales into its so-called "depth-of-repeat" components, i.e., the number of people who have made at least one repeat purchase, the number of people who have made at least two repeat purchases, and so on.

For the purposes of reporting new product sales performance, many research firms (and their clients) have long used the following decomposition of new product sales:

$$
\begin{aligned}
\text{cumulative sales volume} = {} & \text{cumulative \# individuals who have made a trial purchase} \\
& \times \text{average volume per trial purchase} \\
+ {} & \text{cumulative \# individuals who have made a first repeat purchase} \\
& \times \text{average volume per first repeat purchase} \\
+ {} & \text{cumulative \# additional repeat purchases} \\
& \times \text{average volume per additional repeat purchase}
\end{aligned}
$$

(See, for example, Clarke (1984) and Rangan and Bell (1994).) For presentational simplicity, let us assume that only one unit is purchased on each purchase occasion. We can write the decomposition of sales as:

$$S(t) = T(t) + FR(t) + AR(t) \tag{1}$$

where $S(t)$ is the cumulative sales volume up to time t, $T(t)$ is the cumulative number of triers up to time t, $FR(t)$ is the cumulative number of people who have made at least one repeat (first

repeat) purchase by time t, and $AR(t)$ is the total number of additional repeat purchases (i.e., second repeat or higher) up to time t.

Given the entrenched nature of this decomposition among managers in the CPG industry, our approach to the development of a new product sales forecasting model was to create (independent) models for each of the three components of total sales: $T(t)$, $FR(t)$, $AR(t)$. It was felt that such an approach would facilitate the acceptance of the new model within IRI, and aid its communication to clients.

3 Modeling Trial

We model the cumulative number of triers by time t by developing an expression for $P(\text{trial by } t)$, the probability that a randomly-chosen individual has made a trial purchase by time t. For a market comprising N individuals (i.e., the size of the panel), we have

$$T(t) = N \cdot P(\text{trial by } t) \tag{2}$$

We develop our model for $P(\text{trial by } t)$ using a stochastic modeling approach in which we make a set of assumptions about consumer behavior, translate these into probabilistic terms and then derive the complete model. In particular, we make the following four assumptions:

- Only a fraction p_0 of the N individuals in the market will ever make a trial purchase. This accounts for the fact that some individuals will simply not be in the market for the product. For example, one would typically expect that a new brand of cat food will not be purchased by households that do not have a cat.

- For each individual who will eventually make a trial purchase, the time to the trial purchase is treated as a random variable and can be characterized by an exponential distribution with (latent) rate parameter λ_T.

- Beyond the underlying exponential process, marketing activities (e.g., newspaper features, in-store displays, coupons, price changes) will increase or decrease a household's likelihood of making a trial purchase in any given week. Formally, the effects of these marketing ac-

tivities on the distribution of time to trial are incorporated using the proportional hazards framework (e.g., Helsen and Schmittlein 1993; Jain and Vilcassim 1991).)

- Holding the effects of marketing activities constant, households differ in their underlying probability of making a trial purchase by a given point in time after the launch. For example, some households are heavy category buyers and are therefore more likely to make a trial purchase earlier than light category buyers. Formally, λ_T is assumed to be distributed across the population of eventual triers according to a gamma distribution with shape parameter r_T and scale parameter α_T.

Expressing these four assumptions in mathematical terms results in the following expression for $P(\text{trial by } t)$ (Fader, Hardie, and Zeithammer 2003):

$$P(\text{trial by } t) = p_0 \left[1 - \left(\frac{\alpha_T}{\alpha_T + A(t)}\right)^{r_T}\right], \quad t = 1, 2, \ldots \tag{3}$$

where $A(t) = \sum_{i=1}^{t} \exp\left(\boldsymbol{\beta}_T' \boldsymbol{x}(i)\right)$, $\boldsymbol{x}(i)$ denotes the vector of marketing activity variables in week i and $\boldsymbol{\beta}_T$ denotes the impact of these variables on the trial purchase probabilities.

To illustrate this model of trial purchasing, we estimate its parameters using data on the trial purchasing of the 2799 panelists over the first 24 weeks of purchasing data for the Kiwi Bubbles product. We have three different marketing activity variables: one that captures the extent of in-store promotional activity each week, one that captures exposure to manufacturer coupons, and one for the exposure to television advertising. The latter two measures are created as exponentially-smoothed versions of the raw exposure data in order to capture the effects of customer forgetting and wear-out over time.

The six parameter estimates for the trial model are substituted into (3) which, coupled with (2), is then used to forecast trial purchasing to the end of the new product's first year in the test market. (This forecast is conditioned on the actual values of the three marketing activity variables as observed in the two markets over weeks 25–52.) This is plotted in Figure 2 along with the actual level of trial purchasing that occurred over this period.

The performance of the model is quite impressive, with effectively no degradation in its overall tracking capability even as we move far away from the 24-week calibration period. It

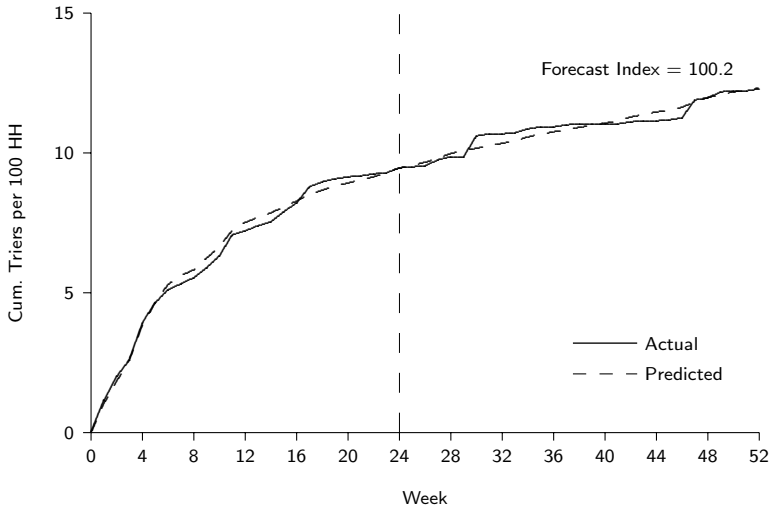

Figure 2: Illustrative Results for Trial Model

is clear that the inclusion of the covariate effects helps capture some of the irregular curvature that exists within the calibration period (e.g., around week 12), yet there are some irregularities in the forecast period that the model misses. It appears to be rather serendipitous that these peaks and valleys cancel each other out to a large extent.

While the degree of accuracy seen here (overforecasting actual 52-week sales by 0.2%) is unusually high, we have seen, in general, that the trial model is consistently strong (generally within 5–10% over a similar timeframe). Detailed evidence about the performance of the trial model can be found in Fader, Hardie, and Zeithammer (2003).

4 Modeling First Repeat

The next step in model development is to derive an expression for $FR(t)$, the cumulative number of people who have made at least one repeat purchase of the new product by time t. In developing this model, as well as our model for additional repeat purchases in the next subsection, we need to make some assumptions about the nature of the data at our disposal. Time is measured on the positive real line, where the origin corresponds to the launch of the new product. By convention, integer values correspond to the end of each week; for example, $t = 2$ corresponds to the end of the second week since the launch of the new product. With consumer panels such as those associated with the BehaviorScan® service, the time of purchase is *recorded* on the real time line (e.g., on a daily or even hourly basis). However, for the purposes of reporting and

model development, time is *processed* in an integer manner (i.e., the data are interval-censored). Eskin (1973, footnote 2, p. 118) notes that when data are processed at a weekly level, a first repeat purchase cannot be made in the same week as its corresponding trial purchase. When such an event does occur in the data, the first repeat purchase must be coded as occuring in the following week. Following this convention, the earliest a jth repeat purchase can occur is in week j+1.

With this data constraint in mind, let us consider how an individual could have made a first repeat purchase of the new product by the end of week 4:

- she could have made a trial purchase in week 1 and made a second purchase (i.e., her first repeat purchase) somewhere in the intervening three weeks,
- she could have made a trial purchase in week 2 and a second purchase sometime in the following two weeks, or
- she could have made a trial purchase in week 3 and her first repeat purchase sometime in the following week.

Letting $P(\text{first repeat by } t \,|\, \text{trial at } t_0)$ denote the probability that a randomly-chosen individual who made a trial purchase at t_0 has made a first repeat purchase by time t ($> t_0$), the above logic implies we can express cumulative first repeat purchases at time t as

$$FR(t) = \sum_{t_0=1}^{t-1} P(\text{first repeat by } t \,|\, \text{trial at } t_0)\big[T(t_0) - T(t_0 - 1)\big] \tag{4}$$

where $T(t_0) - T(t_0 - 1)$ is the number of incremental triers in week t_0. (By definition, $T(0) = 0$.)

We derive an expression for $P(\text{first repeat by } t \,|\, \text{trial at } t_0)$ using exactly the same logic as for the trial purchasing probability $P(\text{trial by } t)$. In other words, we apply a set of assumptions equivalent to those discussed in the previous subsection. Specifically, we assume that: (1) only a fraction p_1 of the triers will ever make a repeat purchase, (2) the time from trial to first repeat purchase (for those triers who will eventually make a first repeat purchase) can be characterized as an exponentially distributed random variable, (3) marketing activities influence first-repeat times according to the proportional hazards framework, and (4) each household's

exponential rate parameter varies in accordance with a gamma distribution. Taken together, these assumptions give us

$$P(\text{first repeat by } t \,|\, \text{trial at } t_0) = p_1 \left[1 - \left(\frac{\alpha_{FR}}{\alpha_{FR} + B(t) - B(t_0)}\right)^{r_{FR}}\right],\ t = t_0+1, t_0+2, \ldots \quad (5)$$

where $B(t) = \sum_{i=1}^{t} \exp\left(\boldsymbol{\beta}'_{FR}\boldsymbol{x}(i)\right)$, $\boldsymbol{x}(i)$ denotes the vector of marketing activity variables in week i and $\boldsymbol{\beta}_{FR}$ denotes the impact of these variables on first repeat purchase probabilities. Note that the structure of this expression mirrors that of $P(\text{trial by } t)$ with the exception that we have $B(t) - B(t_0)$, instead of $B(t)$ alone. $B(t)$ captures the effects of the marketing activities since the launch of the new product; subtracting $B(t_0)$ removes the effects of the marketing activities up to the time of the trial purchase, which should have no direct bearing on purchases beyond trial.

We illustrate this model of first repeat purchasing using the Kiwi Bubbles dataset once again, employing the same three covariates (in-store promotions, coupons, television advertising exposure) as before, but with a different set of coefficients to capture the unique impact that these effects might have on converting triers into first repeaters. The estimated parameters are substituted into (5) which, coupled with (4), is then used to forecast first repeat purchasing to the end of the new product's first year in the test market. (Note that (4) conditions on trial purchasing; forecasts generated using the expression are based on forecast trial sales, as generated using (3) and (2).) This forecast is plotted in Figure 3 along with the actual level of first repeat purchasing that occurred over this period.

In sharp contrast to the results of the trial model, the forecast for first repeat is not especially accurate. While the model tracks the actual data quite well through the calibration period, its projection grows too rapidly throughout the forecast period. The 52-week forecast is 17.2% above the actual year-end level of first-repeat sales.

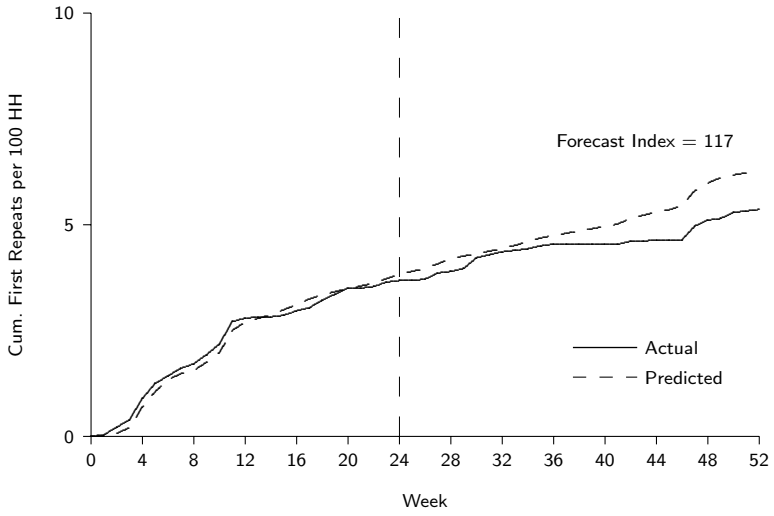

Figure 3: Illustrative Results for First Repeat Model

5 Modeling Additional Repeat

Central to our model for $AR(t)$, the total number of additional repeat purchases by time t is the so-called depth-of-repeat decomposition

$$AR(t) = \sum_{j \geq 2} R_j(t) \tag{6}$$

where $R_j(t)$ is the cumulative number of individuals who have made at least j repeat purchases by time t. But how do we characterize $R_j(t), j = 2, 3, \ldots$?

In the spirit of our first repeat model, let us consider how an individual could have made a second repeat purchase by the end of week 5:

- she could have made her first repeat purchase in week 2 (which implies her trial purchase occurred in week 1) and made a third purchase of the new product (i.e., her second repeat purchase) somewhere in the intervening three weeks,
- she could have made her first repeat purchase in week 3 and a second repeat purchase sometime in the following two weeks, or
- she could have made her first repeat purchase in week 4 and her second repeat purchase sometime in the following week.

Letting $P(\text{second repeat by } t \,|\, \text{first repeat at } t_1)$ denote the probability that a randomly-

chosen individual who made their first repeat purchase at time t_1 has made a second repeat purchase by time t ($> t_1$), the above logic implies we can express the cumulative number of individuals who have made at least their second repeat purchase by time t as

$$R_2(t) = \sum_{t_1=2}^{t-1} P(\text{second repeat by } t \,|\, \text{first repeat at } t_1)\big[FR(t_1) - FR(t_1 - 1)\big]$$

where $FR(t_1) - FR(t_1 - 1)$ is the number of individuals who made their first repeat purchase in week t_1. (By definition, $FR(1) = 0$.)

More generally, we have

$$R_j(t) = \sum_{t_{j-1}=j}^{t-1} \Big\{ P(j\text{th repeat by } t \,|\, (j-1)\text{th repeat at } t_{j-1}) \times \big[R_{j-1}(t_{j-1}) - R_{j-1}(t_{j-1} - 1)\big] \Big\},\ j = 2, 3, \ldots \tag{7}$$

where $P(j\text{th repeat by } t \,|\, (j-1)\text{th repeat at } t_{j-1})$ is the probability that a randomly-chosen individual who made their $(j-1)$th repeat purchase at time t_{j-1} has made a jth repeat purchase by time t ($> t_{j-1}$) and $R_{j-1}(t_{j-1}) - R_{j-1}(t_{j-1} - 1)$ is the number of individuals who made their $(j-1)$th repeat purchase in week t_{j-1}. (By definition, $R_j(t) = 0$ for $t \leq j$.) Clearly, $R_1(t) = FR(t)$.

In order to go from this sales decomposition to a workable model, we need to develop an expression for $P(j\text{th repeat by } t \,|\, (j-1)\text{th repeat at } t_{j-1})$. The approach taken closely resembles that used for trial and first repeat. The problem we face, however, is the sparsity of data as we move to higher levels of data — in most cases we will have little data to calibrate a model for, say, $P(6\text{th repeat by } t \,|\, 5\text{th repeat at } t_5)$, let alone a higher-order quantity such as $P(12\text{th repeat by } t \,|\, 11\text{th repeat at } t_{11})$. The challenge here is that we need to project this depth-of-repeat process to levels of repeat purchasing that we do not observe in our calibration dataset.

There are several different ways we can extrapolate the model parameters to be able to generate a forecast of $AR(t)$. The approach taken in our work draws heavily on Eskin (1973); also see Fader and Hardie (1999) and Kalwani and Silk (1980).

Plotting a number of $P(j\text{th repeat by } t \,|\, (j-1)\text{th repeat at } t_{j-1})$ curves using actual repeat buying data, Eskin observed that the structure of these empirical repeat purchase curves roughly followed the stylized curves given in Figure 4. In particular, he noted that they had three characteristics:

i. The curve for each depth-of-repeat class (j) exhibits an exponential growth pattern leveling off to an asymptote — as is typically associated with a cumulative trial curve.

ii. Across depth-of-repeat classes, the curves are approximately parallel.

iii. The asymptote (i.e., ultimate conversion proportion) increases, at a decreasing rate, as j increases; this means, for example, that the proportion of consumers who have made their jth repeat purchase within 52 weeks of their $(j-1)$th repeat purchase increases with j.

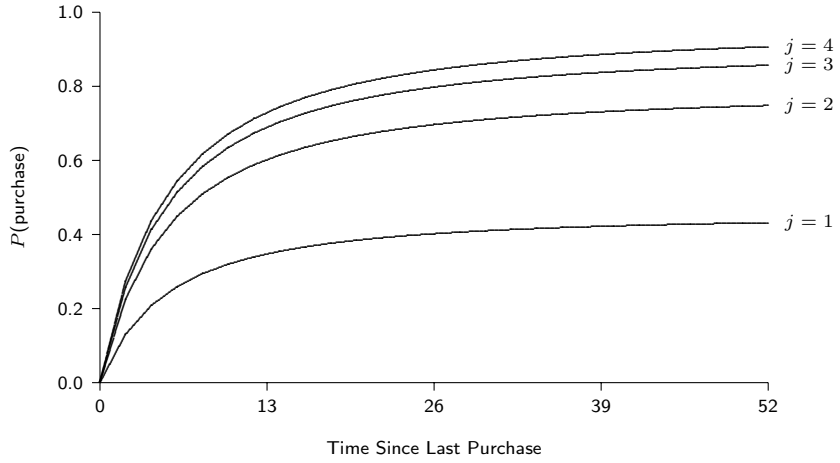

Figure 4: Depth-of-Repeat Curves

We proceed by assuming that all the dynamics across depth-of-repeat can be captured through a structure placed upon p_j, i.e., the fraction of panelists who have made a $(j-1)$th repeat purchase and will eventually make a jth repeat purchase (for $j \geq 2$). We assume that the other parts of the model exhibit no dynamics across depth-of-repeat classes, i.e., the same heterogeneity distribution is assumed to hold across all depth-of-repeat classes (for $j \geq 2$) and the effects of marketing activities on the time to the next purchase are assumed to be invariant over this range as well.

As a consequence of (iii), Eskin proposed an expression for the evolution of the ultimate

conversion proportions equivalent to

$$p_j = p_\infty(1 - e^{-\theta j}),\ j \geq 2 \tag{8}$$

Initial analyses reported in Eskin (1973) indicate that p_∞ is less than 1. However, subsequent experience reported in Eskin and Malec (1976) indicates that $p_\infty \approx 1$, which is consistent with the authors' experience.

Making the same assumptions as for trial and first repeat regarding exponential interpurchase times, the immediate effects of marketing activities, and cross-sectional heterogeneity, we have the following expression for the probability that a randomly chosen panelist who made her $(j-1)$th repeat purchase at t_{j-1} has made a jth repeat purchase by time t $(> t_{j-1})$:

$$P(j\text{th repeat by } t \,|\, (j-1)\text{th repeat at } t_{j-1}) = p_j \left[1 - \left(\frac{\alpha_{AR}}{\alpha_{AR} + C(t) - C(t_{j-1})}\right)^{r_{AR}}\right], \quad t = t_{j-1}+1, t_{j-1}+2, \ldots \tag{9}$$

where $C(t) = \sum_{i=1}^{t} \exp\left(\boldsymbol{\beta}'_{AR}\boldsymbol{x}(i)\right)$, $\boldsymbol{x}(i)$ denotes the vector of marketing activity variables in week i and $\boldsymbol{\beta}_{AR}$ denotes the impact of these variables on additional repeat purchase probabilities.

Because the AR model involves multiple levels of repeat purchasing, the process of generating a forecast is a bit more complicated than for trial and first repeat. Substituting the parameter estimates into (8) and (9), we compute $P(j\text{th repeat by } t \,|\, (j-1)\text{th repeat at } t_{j-1})$ for $j = 2, 3, \ldots$. (For a one-year forecast horizon, we stop this process at $j = 51$.) We then compute the R_j in a forward recursive manner using (7): a forecast of second repeat purchasing is generated conditional on the first repeat sales forecast (not actual), then a forecast of third repeat purchasing is generated conditional on this second repeat forecast, and so on. These are substituted into (6) to give us an overall forecast of additional repeat sales. The corresponding forecast for the Kiwi Bubbles dataset is plotted in Figure 5, along with the actual level of additional repeat purchasing that occurred over this period.

In sharp contrast to Figures 2 and 3, the fitted values for the additional repeat model do not track the actual values very well through the calibration period. This reflects the fact that we are not estimating parameters to fit a single curve, but instead are trying to capture a family

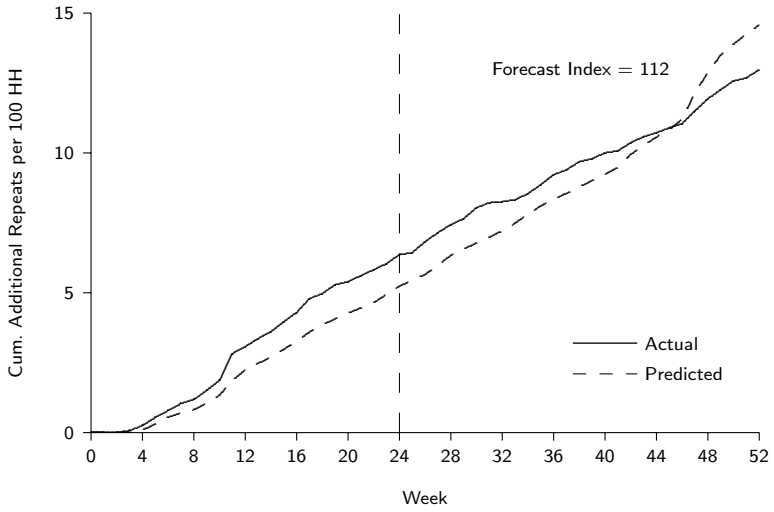

Figure 5: Illustrative Results for AR Model

of curves using a highly parsimonious structure imposed on the AR model parameters. The good news is that there is no degradation whatsoever as the model projection moves away from the 24-week calibration period. The fact that the model over-predicts the actual data by 12% at week 52 largely reflects the fact that these projections are conditioned on the substantial over-forecast seen earlier for the first repeat model. In other words, if we were to layer these AR model results on top of *actual* (not predicted) first repeat sales, the year-end forecast index would be 112/117 = 96.

The ultimate goal of this forecasting procedure is to pull together these three separate forecasts (as noted in (1)) to provide us with an overall sales forecast for the new product. This is illustrated in Figure 6, along with the corresponding actual sales numbers.

This combined graph offers a different — but much more meaningful — perspective on each of the three model components. The over-forecast for first repeat has a relatively small impact on the overall sales decomposition, since FR sales represent a relatively small piece of total sales for Kiwi Bubbles. The same holds for the poor tracking we witnessed for the additional repeat model within the calibration period. At the same time, we begin to see the critical importance of an accurate AR projection. It is evident that virtually all of the sales growth in the forecast period stems from additional repeat, and the extrapolative process at the heart of the AR model seems to perform quite well.

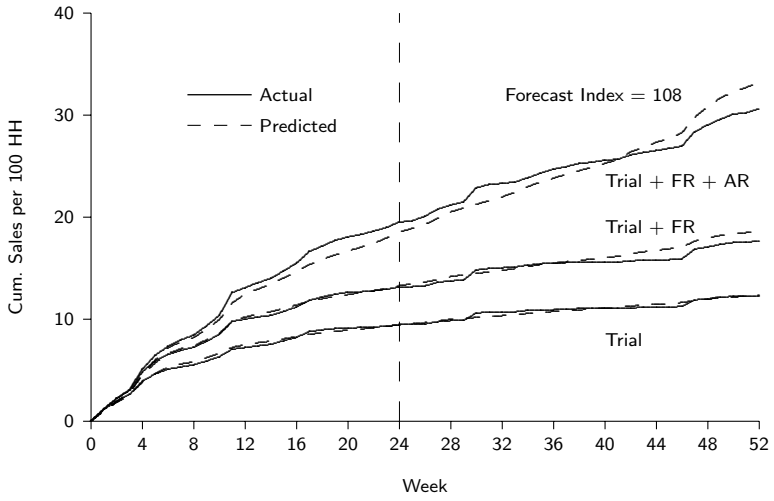

Figure 6: Creating an Overall Sales Forecast

6 Model Validation

The key question regarding any new product sales forecasting model is: can it accurately predict sales volume? Ultimately we are interested in how well the forecasts of national sales match up to the realized sales when the product is finally launched. The challenges associated with this are legion (Urban and Katz 1983). One source of error concerns the projection of the test market results to the national level, controlling for differences between the test markets and the wider market (Berdy 1965; Gold 1964). But this assumes that the within-test-market forecasts are accurate. In this model validation exercise, we examine the accuracy of the within-test-market sales forecasts generated using the model developed above. That is, the model is calibrated on purchasing data for the first 24 weeks of the test market and then a forecast of total sales to the end of the new product's first year in the test market is generated. (As with the above example, this forecast is conditioned on the actual values of the marketing activity variables as observed in the test market over weeks 25–52.) We then compare this forecast to the actual sales observed in the test market.

The challenge faced when undertaking this validation exercise was to find past BehaviorScan® tests that had lasted a whole year. Most tests had a duration of six months; as soon as a sales forecast had been generated using the first 24 weeks of data, the client would typically terminate the test market and make a "go national" decision on the basis of this forecast. As such, most of the data collected using past tests could not be used.

We were able to identify twelve new products (labelled A–L) that underwent year-long BehaviorScan® tests; product J corresponds to the Kiwi Bubbles example above. In Table 1, we report, for each product, the index of year-end forecast accuracy (WK52 Index), computed as 100 × expected week 52 cumulative total sales / actual week 52 cumulative total sales.

Product	WK52 Index	Product	WK52 Index
A	98	G	90
B	81	H	97
C	109	I	99
D	106	J	108
E	111	K	96
F	82	L	101

Table 1: Validation Results

Across these twelve products, the average WK52 index is 98, and there is no indication that these overall forecasts (or their underlying components) show any upward or downward biases. One might argue that this overall mean is a bit misleading since it allows positive and negative forecast errors to cancel out, but even the average *absolute* percent error of 7.7% is well within a tolerable range.

A much tougher validation is to see how well the model performs when we use a shorter calibration period for parameter estimation. Not only does this allow us to scrutinize model performance more carefully, but it also reflects the reality that many new product managers are facing today. A recent article in Advertising Age (2002) discussed a pressing need for managers to obtain accurate forecasts much more quickly than in the past. As one researcher pointed out, "We want to have preventive tools early on. After the fact it's sort of like accident reports."

With these motivations in mind, we re-estimated the model using only 12 weeks of data for a subset of the twelve tests shown above. The overall mean index of 102 is still impressive, although the average absolute percent error expands to 13.5%. This reflects greater variability in these indexed forecasts — they range from a low of 71 to a high of 122. (In contrast none of the 24-week forecasts was off by more than 20%.) It should come as no surprise that the two components of the model that capture repeat sales are much more seriously affected by the shorter calibration period than is the trial component.

This twelve-week analysis points out the need to accumulate these forecasts (and the pa-

rameter estimates) across multiple products in order to properly benchmark the performance of future products (and their associated forecasts). As a firm develops such a database, it can better understand which new products are strokes of luck (good or bad) and which ones are simply following the established norms.

Beyond these aggregate forecasts, the model offers useful diagnostics about the relative sizes and growth rates of the underlying components (as per Figure 6). In addition, the parameter estimates can provide some insight about how the impact of the covariate effects (i.e., marketing activities) varies as we move from trial to first repeat to additional repeat. As a firm begins to accumulate these forecasts and parameter estimates across multiple products, it becomes possible to use this database to properly benchmark the performance and responsiveness of a new product that has just been launched.

7 Discussion and Conclusions

It is sadly ironic that the state of the art in forecasting CPG new product sales has not kept pace with advances in data collection technology and computational power. What makes this irony even more pointed is the fact that the basic frameworks conveyed in older methods (such as those put forth by Fourt and Woodlock (1960) and Eskin 1973)) appear to be just as valid today as when they were first developed decades ago. Unfortunately, as practitioners grapple with incorporating new types and sources of data (e.g., marketing mix variables) they often end up dismantling forecasting systems that embodied all this collected wisdom. The authors' experience with one research firm (Information Resources, Inc.) is fairly typical in this regard; fortunately this firm was able to recover from its past missteps and has now regained a leadership position in its commercial forecasting capabilities. In contrast, most firms do not seem to be enjoying similarly fruitful perspectives these days, as evidenced by the poor forecasting performances discussed in a recent industry survey by Kahn (2002).

One of the primary strengths of the modeling approach we presented here is its broad applicability to many types of new products and services. Not only is it well-suited for the wide array of products that fall under the CPG umbrella, but it should be equally useful for virtually any other product/service that emphasizes repeat purchases (and has a sufficiently short purchase

cycle to enable the observation and modeling of several levels of repeat behavior). Examples outside of the CPG arena include service businesses such as banking and telecommunications, as well as e-commerce buying behavior. In a variety of projects involving repeat-buying behavior from these and other industries, we have seen consistently strong evidence that supports the various components of our modeling approach.

One novel non-CPG example was presented in Fader and Dismukes (2002). A slightly modified version of this model was used to capture and project the depth-of-repeat patterns for a leading online travel company. The objective was to understand the short- and long-run impact of the 9/11 disaster on trial and repeat bookings for the company. The model showed that simple summaries using the observed data alone (e.g., looking at differences in the pre- and post-9/11 bookings) greatly understate the actual impact of the event.

Beyond this type of application/extension of the basic model, there are still some unresolved modeling issues, even in the CPG world, that we have not discussed. One such issue is the notion of *distribution build*. One unique characteristic of the BehaviorScan® electronic test market service is the fact that it ensures complete retail distribution for the product of interest. Representatives of IRI ensure that the product is always available in all retail outlets that they wish to track. In contrast, most new products need to fight for shelf space, and it is often a losing battle. CPG manufacturers pay particular attention to the product's ability to hang on to (and possibly expand) its shelf presence, but they are frequently forced to deal with the consequences of retailer delisting, i.e., removing the product from distribution.

It is easy to bring in retail distribution coverage as another covariate in our proportional hazards model, but distribution serves as more than just an information source or sales incentive like coupons or television ads. Instead, it should be treated as a bottleneck that may prevent the widespread purchasing of a product regardless of how well-promoted it may be. This requires a separate model component, one that is not hard to implement, but is outside the scope of the kinds of datasets that we have used here.

As this modeling framework is extended to incorporate other managerial control variables (and as it is applied to a wider array of products and services), there may be the need for some modifications to the basic equations shown here. But this paradigm should be robust enough to handle these refinements with ease, and it should even perform quite well without them. This

contrasts with the ad hoc regression approaches, described at the start of the paper, that made it difficult to adapt older forecasting models to the changes in data availability that occurred in the 1980s. By focusing on the underlying behavioral process (instead of emphasizing atheoretical "curve-fitting" procedures), we have captured the fundamental trial and repeat process in a general, flexible manner. The resulting improvements in forecasting performance and managerial diagnostics are important rewards that will benefit managers who take the time to build the "right" model in the first place.

References

Advertising Age (2002), "Six Month Check In: Catalina Service to Track New Products," 23 September.

Berdy, Edwin M. (1965), "Testing Test Market Predictions: Commments," *Journal of Marketing Research*, *2* (May), 196–200.

Clarke, Darral G. (1984), "G. D. Searle & Co.: Equal Low-Calorie Sweetener (A)," Harvard Business School Case 9-585-010.

Curry D. J. (1993), *The New Marketing Research Systems*, New York: John Wiley & Sons.

Eskin, Gerald J. (1973), "Dynamic Forecasts of New Product Demand Using a Depth of Repeat Model," *Journal of Marketing Research*, **10** (May), 115–29.

Eskin, Gerald J. and John Malec (1976), "A Model for Estimating Sales Potential Prior to the Test Market," *Proceeding 1976 Fall Educators' Conference*, Series No. 39, Chicago, IL: American Marketing Association, 230–33.

Fader, Peter S., and Ryan Dismukes (2002), "Using Depth-of-Repeat Models to Determine the Impact of 9/11 on Online Travel Sales," presentation delivered at the INFORMS Marketing Science Conference at Edmonton, Canada.

Fader, Peter S. and Bruce G. S. Hardie (1999), "Investigating the Properties of the Eskin/Kalwani & Silk Model of Repeat Buying for New Products," in Lutz Hildebrandt, Dirk Annacker, and Daniel Klapper (eds.), *Marketing and Competition in the Information Age*, Proceedings of the 28th EMAC Conference, May 11–14, Berlin: Humboldt University.

Fader, Peter S, Bruce G. S. Hardie, and Robert Zeithammer (2003), "Forecasting New Product Trial in a Controlled Test Market Environment," *Journal of Forecasting*, forthcoming.

Fourt, Louis A. and Joseph W. Woodlock (1960), "Early Prediction of Market Success for New Grocery Products," *Journal of Marketing*, **25** (October), 31–8.

Gold, Jack A. (1964), "Testing Test Market Predictions," *Journal of Marketing Research*, **1** (August), 8–16.

Helsen, Kristiaan and David C. Schmittlein (1993), "Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models," *Marketing Science*, **12** (Fall), 395–414.

Jain, Dipak C. and Naufel J. Vilcassim (1991), "Investigating Household Purchase Timing Decisions: A Conditional Hazard Function Approach," *Marketing Science*, **10** (Winter), 1–23.

Kahn, Kenneth B. (2002),"An Exploratory Investigation of New Product Forecasting Practices," *Journal of Product Innovation Management*, **19** (March), 133–143.

Kalwani, Manohar and Alvin J. Silk (1980), "Structure of Repeat Buying for New Packaged Goods," *Journal of Marketing Research*, **17** (August), 316–22.

Larson, Erik (1992), *The Naked Consumer*, New York: Penguin Books.

Lodish, Leonard M., Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens (1995), "How T.V. Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments," *Journal of Marketing Research*, **32** (May), 125–139.

Rangan, V. Kasturi and Marie Bell (1994), "Nestlé Refrigerated Foods: Contadina Pasta & Pizza (A)," Harvard Business School Case 9-595-035.

Urban, Glen L. and Gerald M. Katz (1983), "Pre-Test-Market Models: Validation and Managerial Implications," *Journal of Marketing Research*, **20** (August), 221–234.