Journal of Forecasting
J. Forecast. **22**, 391–410 (2003)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/for.869

Forecasting New Product Trial in a Controlled Test Market Environment

PETER S. FADER,[1] BRUCE G. S. HARDIE[2]* AND ROBERT ZEITHAMMER[3]
[1] *The Wharton School, University of Pennsylvania, USA*
[2] *London Business School, UK*
[3] *MIT Sloan School of Management, USA*

ABSTRACT

A number of researchers have developed models that use test market data to generate forecasts of a new product's performance. However, most of these models have ignored the effects of marketing covariates. In this paper we examine what impact these covariates have on a model's forecasting performance and explore whether their presence enables us to reduce the length of the model calibration period (i.e. shorten the duration of the test market). We develop from first principles a set of models that enable us to systematically explore the impact of various model 'components' on forecasting performance. Furthermore, we also explore the impact of the length of the test market on forecasting performance. We find that it is critically important to capture consumer heterogeneity, and that the inclusion of covariate effects can improve forecast accuracy, especially for models calibrated on fewer than 20 weeks of data. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS new product sales forecasting; trial and repeat; test market

INTRODUCTION

Central to the development of many new grocery products, often called 'consumer packaged goods' (CPG), is the use of an in-market test prior to a national launch. A manufacturer undertakes such a test to gain a final read on the new product's potential before deciding whether or not to 'go national' with the new product, as well as to evaluate alternative marketing plans.

Since the pioneering work of Fourt and Woodlock (1960) and Baum and Dennis (1961), a number of market researchers have developed forecasting models that generate a one- to two-year forecast of the new product's performance after, say, six months in the test market. The ability to shorten the duration of a market test reduces the cost of the test itself, the risk of a competitor getting to market first (or of being fast follower), as well as the opportunity costs of not going national earlier (assuming the test would result in a 'go national' decision).

* Correspondence to: Bruce G. S. Hardie, London Business School, Regent's Park, London NW1 4SA UK
E-mail: bhardie@london.edu

The vast majority of these forecasting models were developed in the 1960s and 1970s, during which time the gathering of weekly data on in-store merchandising activity (e.g. feature and/or display promotions) was non-existent, unless collected on a custom-audit basis. Consequently most of the models developed in this era did not include the effects of marketing decision variables; two rare exceptions are Eskin (1974) and Nakanishi (1973). With the widespread adoption of the Universal Product Code (UPC) and associated laser scanner technology, information on in-store marketing activity is now readily available.

A key model specification question we answer in this study is whether or not the incorporation of covariates such as marketing decision variables improves a model's forecasting performance. Within the diffusion modelling literature, it has been observed that the omission of covariates rarely has an impact of the accuracy of the forecasts generated by the model (e.g. Bass *et al*., 1994; Bottomley and Fildes, 1998). However, consumer durables—the typical class of product to which such models are applied—are quite different from CPG products and we cannot automatically assume that this result will hold in a different context.[1]

When examining the performance of new CPG products, it is standard practice to separate total sales into trial (i.e. first purchase) and repeat (i.e. subsequent purchases) components; repeat sales are then decomposed into first repeat, second repeat, third repeat (and so on) components. Within the literature on test-market forecasting models, there is a long tradition of building separate models for trial, first repeat, second repeat, etc., and then combining the output from each sub-model to arrive at an overall sales forecast for the new product—see, for example, Eskin (1973) and Fader and Hardie (1999). Although the model parameters may vary from one level of depth-of-repeat to another, the general structure of the model is usually assumed to be the same across each purchasing level.[2] For this reason, our examination of the role of covariates in a model's forecasting performance will focus exclusively on models for the trial component of new product sales, so as to simplify the process of gaining tangible insights. (Such an approach was also used by Hardie *et al*., 1998.)

As we develop models of trial purchasing that incorporate the effects of marketing covariates, we could follow the approach taken in the diffusion modelling literature for a number of years in which covariate effects were 'tacked-on' to existing models of the diffusion process. The problem with this is that it has resulted in some rather *ad-hoc* model specifications (Bass *et al*., 2000). The approach we will take is to start with a clean slate, building new models in which the effects of covariates are explicitly accounted for at the level of the individual customer. The models developed in this way nest most of the established (no-covariate) models of trial purchasing that were examined by Hardie *et al*. (1998).

When we consider the implementation of these models, a question arises concerning the length of the model calibration period. As time pressures continually intensify in many organizations, management no longer has the luxury to wait for the completion of the test market before making further decisions about product rollout—they need to be able to project consumer demand using as little data as possible. However, we would expect there to be a trade-off between the length of

[1] This caution is reinforced by the observation that the 'Bass model', a common diffusion model in the marketing literature, performs poorly in both describing and forecasting the trial sales of new CPG products (Hardie *et al*., 1998).

[2] A problem with such this depth-of-repeat approach is that it can result in misleading inferences about buyer behaviour, as the model formulations fail to recognize any dependence across purchases at the individual-level (e.g. Gupta and Morrison, 1991). While academic researchers have developed models that address this issue, market research companies continue to use this established modelling framework for a number of practical reasons, principally the quality of its forecasts (Fader and Hardie, 1999).

the model calibration period and the model's forecasting performance. This was briefly explored by Hardie *et al*. (1998), who compared the forecasting performance of trial purchase models calibrated using the first 13 weeks versus the first 26 weeks of test market data. In this paper we wish to explore systematically the effects of the length of the test market on forecasting performance.

The paper proceeds as follows. We start by exploring the general structure of a model of trial purchasing, which leads to the identification of eight candidate models. These models are then calibrated on a set of datasets and their forecasting performance computed. We then examine the impact of model structure, along with the length of the model calibration period, on forecasting performance. Finally, we conclude with a discussion of a number of issues that arise from our study, and identify several areas for follow-on research.

MODEL DEVELOPMENT

In building a trial purchase model, we specify a parametric form for $F(t)$, the new product's penetration at time t where the zero point of the time scale corresponds to the time of the introduction of the new product. Once the model parameters have been estimated using data on trial purchases for a given calibration period, the new product's penetration can be forecast out into the future by simply computing the value of $F(t)$ for subsequent values of t. Trial sales estimates for the panel are calculated by multiplying the penetration numbers by the panel size and average trial purchase volume. Market-level trial sales estimates can then be computed by multiplying the panel-level numbers by panel projection factors that account for differences in the mix of households in the panel versus the market as a whole.

Some early attempts at specifying $F(t)$ took a 'curve fitting' approach in which the researchers proposed a flexible functional form designed to 'best fit' the observed data. An example of this is the well-known Fourt and Woodlock (1960) model. In examining numerous cumulative trial curves, they noted that (i) successive increments in cumulative trial declined, and (ii) the cumulative curve approached a penetration limit of less than 100% of the households in the panel. They proposed that incremental trial (i.e. $F(i) - F(i-1)$) be modelled as $rx(1-r)^{i-1}$, where x = the ceiling of cumulative trial (i.e. the penetration limit), r = the rate of penetration of the untapped potential, and i is the number of (equally spaced) time periods since the launch of the new product.

Since then, most developers of new product trial forecasting models have developed models using a stochastic modelling approach in which they make a set of assumptions about consumer behaviour, translate these into probabilistic terms and then derive the complete model. Following Hardie *et al*.'s (1998) classification of the published trial models, the general form of a trial purchase model follows the mixture model specification

$$F(t) = p\int F(t|\theta)g(\theta)d\theta \tag{1}$$

The three 'components' of this general model form, $F(t|\theta)$, $g(\theta)$, and p, are defined as follows:

(1) At the heart of any trial model is the specification of $F(t|\theta)$, the cumulative distribution function (cdf) of an individual panelist's time-to-trial. (This is called the *structural model*.)
(2) When specifying the structural model, we take the perspective of a single panellist. As we move from one panellist to all panellists in the market, we could make the assumption that they are

all perfectly homogeneous. However, heterogeneity is central to marketing thinking—some consumers may be inherently fast buyers while others may be inherently slow buyers. Such household differences can be accommodated by specifying a *mixing distribution* $g(\theta)$ for the structural model's parameters.

(3) In examining data from numerous new product launches, Fourt and Woodlock (1960) observed that the cumulative trial curve almost always approached a penetration limit of less than 100% of the households in the panel. Consequently, they proposed that a trial model should incorporate a ceiling on cumulative trial via a penetration limit term. The inclusion of such a term is quite plausible as, in most situations, some people will never be in the market for the new product no matter how long they wait. For example, one would typically expect that diapers will not be purchased by panellists who do not have children (or grandchildren) under 3 years old. This implies that the assumed cdf for a panellist's time-to-trial only applies to those panellists that will eventually try the new product, and that the associated probabilities are therefore conditional. The probability of a randomly chosen individual eventually trying the new product is p, which can also be interpreted as the proportion of the market that will try the new product. Although $1 - p$ represents the proportion of the market that will never try the new product, the penetration limit p is sometimes called the 'never triers' term.

Specific models follow by defining the various model components. For example, the trial model at the heart of Massy's (1969) STEAM model assumes $F(t|\theta)$ is Weibull, $g(\theta)$ is gamma for the rate parameter of Weibull distribution, and $p \leq 1$. A model proposed by Anscombe assumes $F(t|\theta)$ is exponential, $p = 1$ (i.e. no 'never triers' term), and $g(\theta)$ is gamma. The continuous time equivalent of the Fourt and Woodlock model assumes $F(t|\theta)$ is exponential, $g(\theta)$ puts unit mass on $\theta = \lambda$, and $p \leq 1$ (Anscombe, 1961). Herniter (1971) assumed an Erlang-k structural model with an exponential mixing distribution. It must be noted that no model developer has provided direct evidence for his choice of structural model; the particular distributions employed have simply been *assumed* to be correct.

The logical point at which to incorporate the effects of marketing mix variables is at the level of the individual, i.e. via $F(t|\theta)$. (This is in contrast to the approach initially taken in the diffusion modelling literature in which the covariate effects were incorporated directly into the aggregate-level function, $F(t)$.) Today, the standard approach for incorporating the effects of covariates into event-time models is the proportional hazard approach. (See Appendix A for a brief review.) This leads to $F(t|\theta, \mathbf{X}(t), \boldsymbol{\beta})$, an individual-level with-covariates cdf for the distribution of time-to-trial, where $\mathbf{X}(t)$ represents the covariate path up to time t and $\boldsymbol{\beta}$ denotes the effects of these covariates. Drawing on the general mixture model specification given in (1), we can therefore write the general form of a with-covariates trial purchase model as

$$F(t|\mathbf{X}(t), \boldsymbol{\beta}) = p\int F(t|\theta, \mathbf{X}(t), \boldsymbol{\beta})g(\theta)d\theta \tag{2}$$

In order to move from generalities to a specific model of trial purchasing, we must make decisions about the nature of $F(t|\theta)$ and $F(t|\theta, \mathbf{X}(t),\boldsymbol{\beta})$, $\mathbf{g}(\theta)$, and p. As previously noted, model developers have assumed a particular specification for $F(t|\theta)$ without providing direct evidence for their choice of structural model. In Appendix B (available at http://brucehardie.com/papers/fhz_appendix_b.pdf), we report on an analysis in which we conclude that the exponential distribution is the 'correct' structural model for trial purchasing. This implies that $F(t|\theta) = 1 - \exp(-\theta t)$ and therefore $F(t|\theta, \mathbf{X}(t),\boldsymbol{\beta}) = 1 - \exp(-\theta\, A(t))$ where $A(t) = \sum_{i=1}^{\text{Int}(t)} \exp[\boldsymbol{\beta}'\boldsymbol{x}(i)] + [t - \text{Int}(t)]\exp[\boldsymbol{\beta}'\boldsymbol{x}$

$(\text{Int}(t+1))]$, with $\boldsymbol{x}(i)$ denoting the vector of covariates for time period i. (See Appendix A for derivations.)

Having specified the underlying structural model, equations (1) and (2) suggest that a trial forecasting model can be characterized in terms of three 'components': (i) whether or not the existence of a group of 'never triers' is explicitly acknowledged, (ii) whether or not heterogeneity in consumer buying rates is explicitly modelled, and (iii) whether or not the effects of marketing decisions variables are incorporated. The exclusion of a 'never triers' component corresponds to constraining p to 1.0. Not including the effects of unobserved heterogeneity corresponds to specifying $g(\theta)$ such that we have a point mass on $\theta = \lambda$. To accommodate the effects of unobserved heterogeneity, we will assume that the latent trial rate θ is distributed according to a gamma mixing distribution; i.e

$$g(\theta \mid r,\alpha) = \frac{\alpha^r \theta^{r-1} e^{-\alpha\theta}}{\Gamma(r)}$$

where r and α are, respectively, the shape and scale parameters.

Looking at all possible combinations of these components (i.e. presence/absence of penetration limit, heterogeneity, and covariates) gives us eight candidate models. The equations for the eight models corresponding to the inclusion/exclusion of each of these three model components can be obtained by evaluating equations (1) and (2), and are presented in Table I. This table also presents the naming convention we will use to label these eight models for the remainder of the paper. All eight models feature an exponential structural model, and thus begin with the letter 'E'. The four models that have gamma heterogeneity are called 'EG'. Several of the models are suffixed with a 'N' and/or 'C' to describe the presence of a 'never triers'/penetration limit term and/or covariates. Thus the simplest model, the one parameter pure exponential, is simply known as E while the most complex model, EG_NC, encompasses all three components.

Table I. Functional forms for candidate trial models

Model	Functional form	'Never triers"	Heterogeneity	Covariates
E	$F(t) = 1 - e^{-\lambda t}$	N	N	N
E_N	$F(t) = p\left[1 - e^{-\lambda t}\right]$	Y	N	N
EG	$F(t) = 1 - \left(\frac{\alpha}{\alpha + t}\right)^r$	N	Y	N
EG_N	$F(t) = p\left[1 - \left(\frac{\alpha}{\alpha + t}\right)^r\right]$	Y	Y	N
E_C	$F(t) = 1 - e^{-\lambda A(t)}$	N	N	Y
E_NC	$F(t) = p\left[1 - e^{-\lambda A(t)}\right]$	Y	N	Y
EG_C	$F(t) = 1 - \left(\frac{\alpha}{\alpha + A(t)}\right)^r$	N	Y	Y
EG_NC	$F(t) = p\left[1 - \left(\frac{\alpha}{\alpha + A(t)}\right)^r\right]$	Y	Y	Y

These models will be calibrated on a set of datasets and their forecasting performance computed. We then determine whether any systematic patterns in each model's forecasting performance can be linked to its components. Additionally, we will examine the impact of the length of the model calibration period on forecasting performance, along with any interactions between model formulation and calibration period length.

EMPIRICAL ANALYSIS

The data used in this study come from market tests conducted using Information Resources, Inc.'s (IRI) *BehaviorScan* service. *BehaviorScan* is a controlled test-marketing system with consumer panels operating in several markets, geographically dispersed across the USA. At the time these data were collected, there were eight active markets: six of these were targetable TV markets (Pittsfield, MA, Marion, IN, Eau Claire, WI, Midland, TX, Grand Junction, CO, and Cedar Rapids, IA), the other two were non-targetable TV markets (Visalia, CA and Rome, GA). (See Curry, 1993 for further details of the *BehaviorScan* service.) We have five datasets (labelled A–E), each associated with a new product test (lasting one year) conducted in one of the targetable TV markets between 1989 and 1996. The tested products are from the following categories: shelf-stable (ready-to-drink) juices, cookies, salty snacks, and salad dressings.

The recorded individual panellist trial times are interval-censored; that is, the week of trial purchase is reported. We therefore create a dataset containing 52 weekly observations, each observation being the number of panellists who tried the new product during the week in question. Additionally we have information on the marketing activity for the new product over the 52 weeks the new product was in the test market. For four of the datasets (A–D), this comprises a standard scanner data measure of promotional activity (i.e. any feature and/or display), along with measures of advertising and coupon activity. To account for carryover effects, the advertising and coupon measures are expressed as standard exponentially-smoothed 'stock' variables (e.g. Broadbent, 1984). No advertising data were available for the fifth dataset (E); however, an additional promotional tool, an instantly redeemable coupon, was used and this was captured via a dummy variable.

The model parameters are estimated using the method of maximum likelihood. Given the interval-censored nature of the data, the general log-likelihood function is given by

$$LL = \sum_{i=1}^{t_c} n_i \ln[F(i) - F(i-1)] + \left(N - \sum_{i=1}^{t_c} n_i\right) \ln[1 - F(t_c)]$$

where n_i is the number of triers in week i, N is the number of households in the panel, and t_c is the number of weeks of data used for model calibration. (The exact equation is derived by substituting in the specific expression for $F(t)$ from Table I.) Using standard numerical optimization software, we find the values of the model parameters that maximize this log-likelihood function; these are the maximum likelihood estimates of the model parameters.

For each model × dataset combination, we calibrate the model parameters using the first t_c weeks of data. In order to examine the impact of calibration period length on forecast performance, we vary t_c from 8 to 51 weeks in one-week increments. Using the parameters estimated on the first t_c weeks of data, each model is used to forecast cumulative trial for each of the remaining $(52 - t_c)$ weeks. In summarizing a model's ability to forecast cumulative trial, we are interested in both the week-by-week accuracy and year-end (i.e. week 52) cumulative trial. The appropriate error measures will

be computed for each of the 1760 model specification × calibration period × dataset (8 × 44 × 5) combinations. These will then be analysed to identify the impact of the various model components and calibration period lengths on forecasting performance.

The issue of what error measure(s) a researcher should use to identify the most accurate forecasting method has received much attention in the forecasting literature. One class of measures focuses directly on forecast error; for example, mean absolute error (MAE) and mean-squared error (MSE). However, such measures are scale dependent and therefore cannot be used in comparing models across data series which differ in magnitude. (Our data series differ considerably in magnitude, with 52-week penetration varying from just over 6% to almost 40%.) We therefore consider relative error measures, which remove such scale effects. One measure that is widely used is Mean Absolute Percentage Error (MAPE). While there are subtle theoretical advantages associated with the use of alternative measures—see Armstrong and Collopy (1992), Fildes (1992), and related commentary—MAPE has the advantages of not only being very interpretable but also very appropriate in planning and budgeting situations (Makridakis, 1993). We will therefore focus primarily on the MAPEs calculated over the forecast period for each of the model × calibration period × dataset combinations. (Alternative measures of model performance were computed. For example, we also examined point estimates of forecasting accuracy by computing the percentage error for week 52 alone. However, the results of this analysis parallel the MAPE results to a very high degree; as such we only report the MAPE findings.)

RESULTS

From an applied perspective, our primary interest is in the forecasting performance of each model. However, in order to understand the impact of the length of the calibration period, we will also consider the issue of parameter stability—the extent to which a model has calibration period length-invariant parameter estimates. In our search for the best model(s), we therefore examine both dimensions of model performance—forecasting ability and parameter variation. We discuss each performance dimension separately, but we show that their results interact significantly. Together these criteria will jointly help us to identify the most appropriate and important characteristics for a model of trial purchase behaviour.

Analysis of forecasting results

We begin with an examination of the MAPE results, which are summarized in Figure 1. Each point in the graph represents the average MAPE across all five datasets for each model type and calibration period. For instance, the pure exponential model, represented as an unadorned dashed line, has an average MAPE just over 180% when eight weeks of calibration data are used to forecast sales for the remaining 44 weeks in each dataset. When 28 weeks of calibration data are used, its average MAPE is much improved (but still very poor) at just below 60%.

Several noteworthy patterns are immediately evident. First is the observation that the pure exponential model with no covariates forecasts far worse than the other seven models, regardless of the amount of calibration data available to it. Even when this model uses data from the first 51 weeks to make a forecast for week 52 alone, its resulting absolute percentage error across the five datasets (16%) is still worse than that of several models with utilizing only 12 weeks of calibration data. Thus while simplicity may be a virtue, the pure exponential model is clearly far too oversimplified

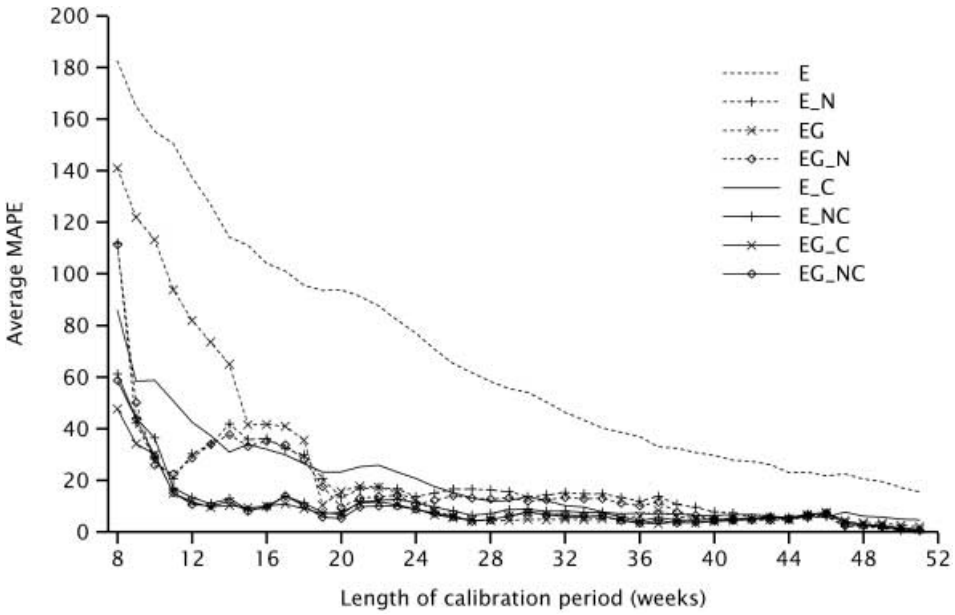

Figure 1. Forecasting errors: all models (average across five datasets)

to be of any value. Because of the very poor forecasts produced by this model, we omit it from all further analyses in this section.

Second, we see that, by week 20, all seven of the remaining models have achieved reasonable levels of MAPE, although there are substantial differences among the models. There appears to be less of an improvement in the model forecasts beyond this point. This observation has important implications for the crucial managerial decision about whether to wait for additional market data before making a final forecast versus making a 'definitive' trial forecast now and sticking with it.

As we examine Figure 1 more carefully, it is evident that three of the models appear to reach their 'elbow' points far earlier than the other models, roughly around week 12. Upon closer inspection, one may notice that all three of these models include covariates as well as any combination of heterogeneity and/or 'never triers' (i.e. E_NC, EG_C, and EG_NC, using the notation from Table I). Not only do each of these three models reach an elbow point faster, but they maintain a slight forecasting advantage over the other models all the way past 30 weeks of calibration data.

Therefore the two principal conclusions we can draw from Figure 1 are that: (a) the inclusion of covariates, if at all possible, is the first critical step in building a 'good' forecasting model, especially when one wishes to use a relatively short calibration period; and (b) the best forecasting models add in at least one other component (heterogeneity and/or 'never triers') with the covariates.

While these are believable and useful findings, they can be refined even further. So as to take a closer look at the interplay among the various components and calibration periods, we present in Figure 2 the forecasting performance results for the four models that include covariates. To make the graph as clear as possible, we only show the forecasts from models calibrated with at least 10 weeks of data.

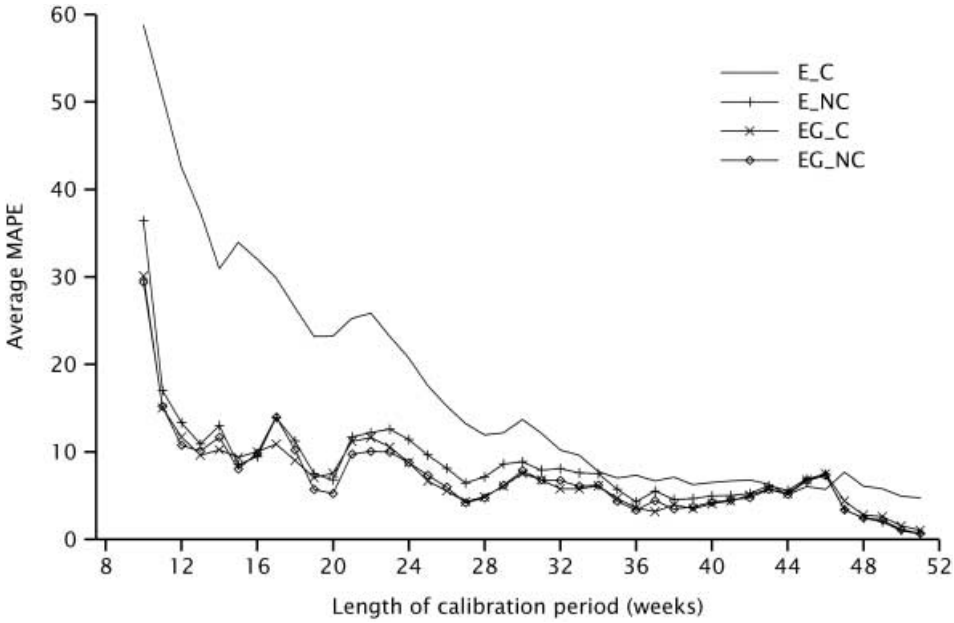

Figure 2. Forecasting errors: models with covariates only

The E_C model (solid line) is clearly inadequate, as suggested by the preceding discussion. At first, the other three models may appear to be essentially indistinguishable from each other, but upon closer inspection it can be seen that the covariate model with 'never triers' only (i.e. E_NC) is consistently less accurate than either or both of the other two models all the way through 40 calibration weeks. The inference to be made here is that while the 'never triers' component appears to help somewhat, it is more important to directly capture heterogeneity in trial rates.

We are left with two strong models in Figure 2, EG_NC and EG_C, with virtually identical forecasting capabilities. While this may appear to be a difficult choice, we favour the EG_C model for several reasons. We can now appeal to its simpler structure, with one less parameter but essentially no loss in forecasting power. The gamma distribution is highly flexible and can accommodate 'never triers' by treating them as 'very slow but eventual' buyers who will enter the market at a late stage (perhaps on the order of years) which, for the standard forecasting horizon, is equivalent to never trying. Furthermore, as we will see later, the parameter estimates associated with the EG_NC specification are often highly unstable, especially when relatively few calibration weeks are available to estimate the model.

This reasoning allows us to declare EG_C as the overall 'winner', and its performance is quite strong indeed, with forecasts generally within 10% of actual even with as few as 12 weeks of calibration data. But an important question remains to be addressed: which model(s) are most suitable when covariates are not available to the analyst? Despite the advances made possible by today's scanning technology, it is easy to conceive of situations in which covariate information (e.g. coupons, advertising, or in-store promotions) may be missing or subject to a variety of measurement errors. It is therefore imperative that we identify a robust model that can produce accurate forecasts without using any such covariates.

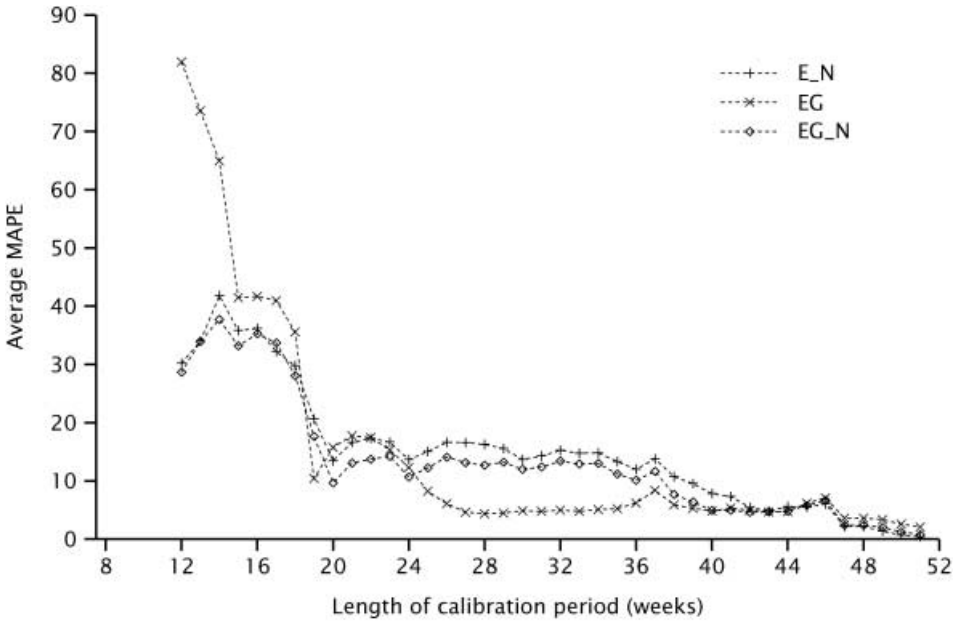

Figure 3. Forecasting errors: models without covariates

In Figure 3 we examine the performance of the three candidate models that ignore covariates (again, we omit the pure exponential model). To enhance interpretability of this graph, we only consider models with at least 12 weeks of calibration data. The results are quite interesting. When relatively few (<18) weeks of calibration data are used, the plain exponential-gamma (EG) model is very poor. (Even in Figure 1 it is clear that this is the second-worst model overall through 18 calibration weeks.) The explanation here is that the EG model is mistaking the unexplained covariate effects strictly as evidence of consumer heterogeneity, and is inferring a very distorted distribution of purchase rates across households. While we observe a slight improvement when "never triers" are allowed to enter the picture (i.e. the EG_N model), the MAPE numbers are probably still too high for the forecasts to be of use from a managerial perspective.

In contrast, however, as the length of the calibration period moves beyond 20 weeks, the simple EG model dramatically improves and becomes the best forecasting model, all the way through 35 weeks of calibration data. Apparently, as the set of consumers entering the market becomes sufficiently diverse, true heterogeneity effects dominate any apparent differences due to early marketing activity, and the underlying gamma distribution is specified more properly.

The conclusions from Figure 3 are as twofold: first, in the absence of covariate effects, extreme caution should be used in making any forecasts with fewer than 20 weeks of calibration data; beyond this point, the EG model appears to be the best choice. While the EG_N model eventually catches up and even surpasses EG (with 40 or more weeks of calibration data) the same arguments as before still apply: the 'never triers' term can be redundant when heterogeneity is explicitly modelled, and the forthcoming parameter stability analysis will clearly show why we favour the model with one component over both.

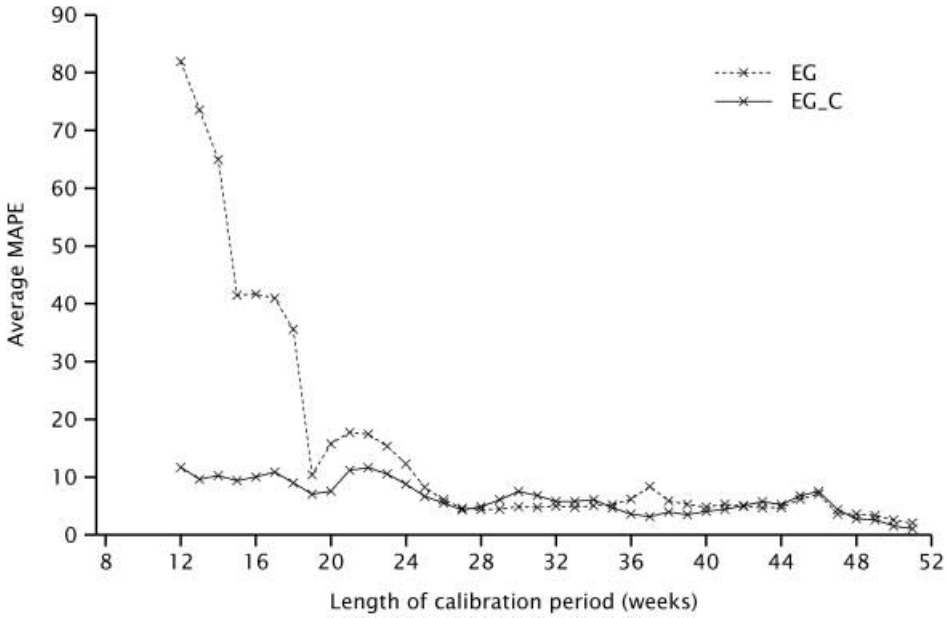

Figure 4. Forecasting errors: exponential-gamma models

To complete our picture of the best models, we compare the forecasting performance of the two EG specifications (with and without covariates) in Figure 4. After the plain EG model catches up with EG_C (with 26 or more weeks of calibration data), the two models are very hard to distinguish from one another. The inclusion of covariate effects has surprisingly little impact in these later weeks (even though several of the datasets have significant promotional activities taking place during this period). At the same time, however, the added complexity from including the covariate parameters appears to cause no harm either. One might have guessed, *a priori*, that the sales impact of promotional activities would be less in later weeks compared to the early weeks of a new product introduction. Under this hypothesis, the static (non-time-varying) nature of the $\boldsymbol{\beta}$ coefficients would lead to systematic overpredictions towards week 52. Apparently, there is no evidence to support such a view. The EG_C model is successfully able to sort out heterogeneity and covariate influences equally well for all calibration periods with 12 or more weeks of data.

To summarize, the exponential-gamma model (with no provision for 'never triers') is highly robust and accurate. If an analyst has proper covariate measures available, the EG_C model appears to offer reliable forecasts starting around week 12. If covariate effects are unavailable or in any way untrustworthy, then the simpler EG model can be employed around week 20. For typical forecasting situations, which often use 26 weeks of calibration data to make forecasts for a 52-week time horizon, the two models are very similar. Other criteria, such as the potential diagnostic value of measuring covariate effects, would play a larger role in model selection.

Analysis of parameter stability

In order to gain insight as to why the forecasting performance of some models is relatively insensitive to the length of calibration period—when compared to other models—we explore the issue of

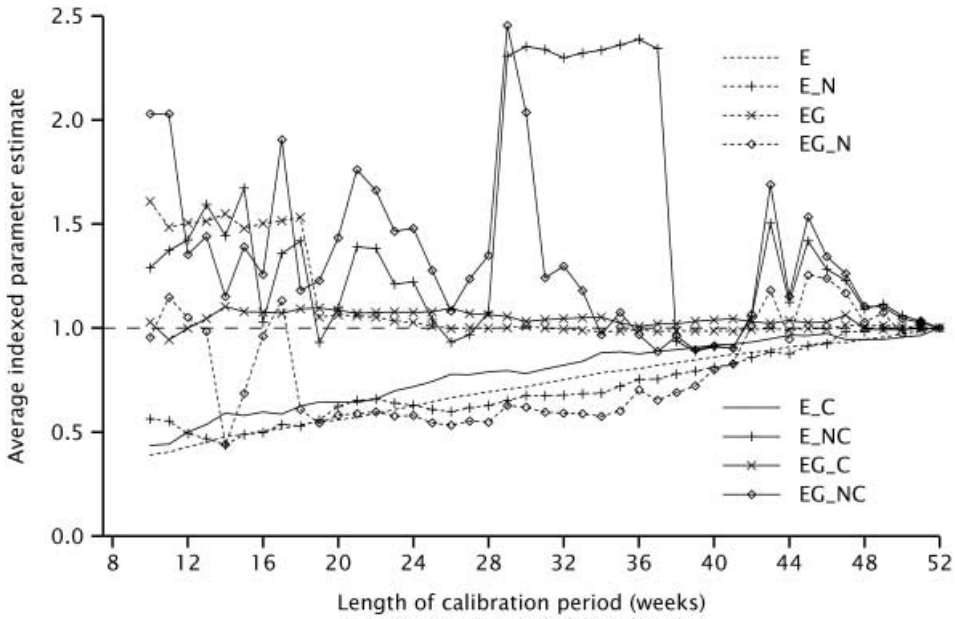

Figure 5. Parameter stability: mean time-to-purchase parameter

parameter stability. In particular, do we see much variation in the parameter estimates as we increase the length of the calibration period (i.e. provide more 'information' for parameter estimation purposes)? If the parameter estimates for a given model specification are relatively insensitive to calibration period length, we would expect to see little variation in forecasting performance as the calibration period changes.

To analyse parameter variation across different model specifications and calibration periods, we created indexed parameter estimates by dividing all parameters for each of the 1760 model × calibration period × dataset combinations by their respective estimates based on 52 weeks of calibration data. This gives us the best possible indication of the loss of information that results from using a shorter (i.e. <52 weeks) calibration period. The across-dataset averages of these indexed parameter values are then plotted for each model specification and calibration period. This approach allows for detection of both systematic biases and random instabilities of the parameters. We first discuss the stability analysis for each of the key parameters and conclude this section by integrating these stability analysis results with our forecasting conclusions.

Probably the most important parameter common to all of our models is the mean of the implied time-to-purchase distribution. This is simply the scale parameter (λ) for the exponential model and the shape parameter (r) divided by the scale parameter (α) for the exponential-gamma model. Figure 5 demonstrates how the estimates of these means vary across model specifications and calibration periods.

This and the subsequent two figures can be read as follows: a 'perfect' model would have indexed parameter values equal to 1.0 for all calibration periods. In other words, such a model would always provide the correct 'full information' (i.e. 52-week) estimate for the parameter of interest, regard-

less of calibration period length. It follows that indexed values above 1.0 indicate overestimates and values below 1.0 are underestimates compared to the 52-week numbers.

Perhaps the most noticeable aspect of this graph is the high degree of instability evident for the three models that involve 'never triers' and at least one other component (i.e. E_NC, EG_N, and EG_NC). These jumps are severe and unpredictable, even for long calibration periods. This is clear evidence of the inadequacies of the 'never triers' component, and a strong indication that using more bells and whistles does not necessarily lead to a better model.

In contrast, the 'winners' here are the same two EG models discussed in the previous subsection—EG and EG_C—which capture the full-information estimates of the mean time-to-trial parameters far better than the other models. The EG_C model is accurate right from week 8, and varies very little over longer calibration periods. As expected, given its tendency to over-forecast with limited calibration data, the pure EG model dramatically overstates the mean over short calibration periods (since it improperly accounts for accelerated purchases due to promotional activity), but settles down quite nicely by week 20. In fact, for most of the longer calibration periods, the EG model is slightly better than EG_C, although both are excellent.

Three of the models without heterogeneity show systematic and highly consistent underestimates of the mean. Even after the forecasts have begun to stabilize for several of these models (e.g. around week 20 for E_N), the underlying parameters are still fairly biased.

For brevity, we skip the stability analysis for the 'never triers' parameter. As just discussed, most of these plots show high degrees of instability. Furthermore, in many cases (especially when heterogeneity is included in the model), the 'never triers' parameter is not significantly different from 1.0 for all calibration-period lengths and so it is not very meaningful to examine its stability.

While the mean time-to-trial parameter demonstrates very different stability patterns across the various model specifications, the covariate parameters do not exhibit such differences in stability performance. Therefore, they do not discriminate effectively among the different models. Nevertheless, their behaviour offers insight about the minimum necessary calibration period, because the parameter estimates are quite erratic up to about week 20 and then settle down to be relatively stable (see Figure 6). We only present the stability plot for the promotion parameter because the plots for other parameters are very similar.

The last parameter remaining to be examined is the r parameter, which reflects the variance of the mixing gamma distribution in the heterogeneous models.[3] By now it should come as no surprise that the models that include both heterogeneity and 'never triers' will be highly unstable, so we omit these two models and only show the stability pattern the exponential-gamma models with and without covariates only (see Figure 7).

As may be expected, the EG_C model performs quite consistently across different calibration periods, while the pure EG model is very unstable until week 19. As discussed earlier, the EG model infers that the underlying heterogeneity distribution is a lot tighter (i.e. lower variance) than is actually the case, since it is tricked by the large number of early, promotion-induced buyers. Even with as many as 18 weeks of calibration data, the average value of the r parameter is about 50 times larger than its 52-week estimate.

After that point, however, the graph shows the largest contrast we have seen between these two models. The estimated values of r for the EG model move almost precisely to the 52-week values,

[3] One summary measure of heterogeneity in probability models is the coefficient of variation of the mixing distribution (Morrison and Schmittlein, 1988). For the gamma distribution, C.V. = $1/\sqrt{r}$. Consequently, we can interpret r as a measure of heterogeneity.

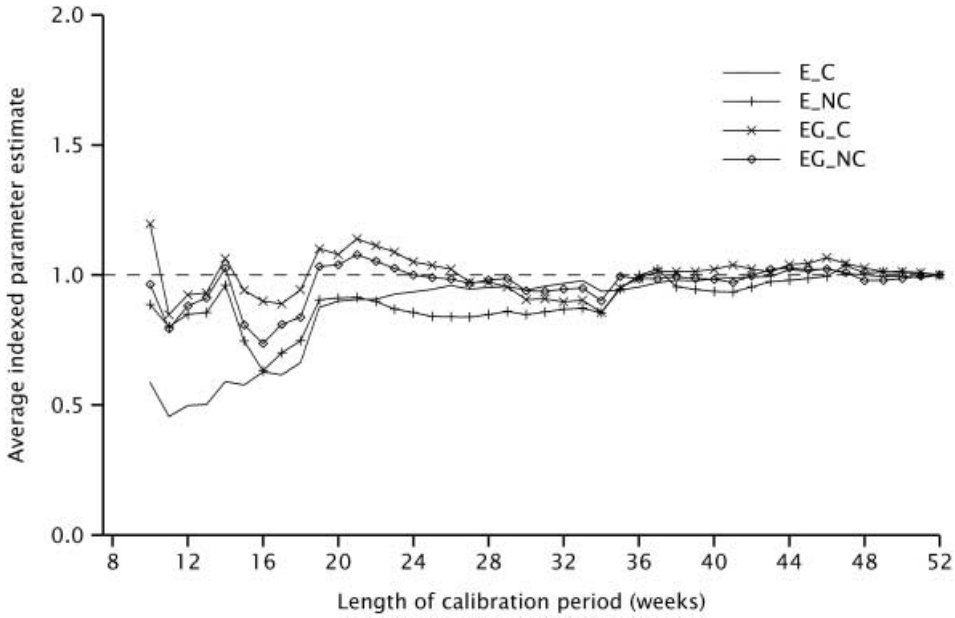

Figure 6. Parameter stability: promotion parameter

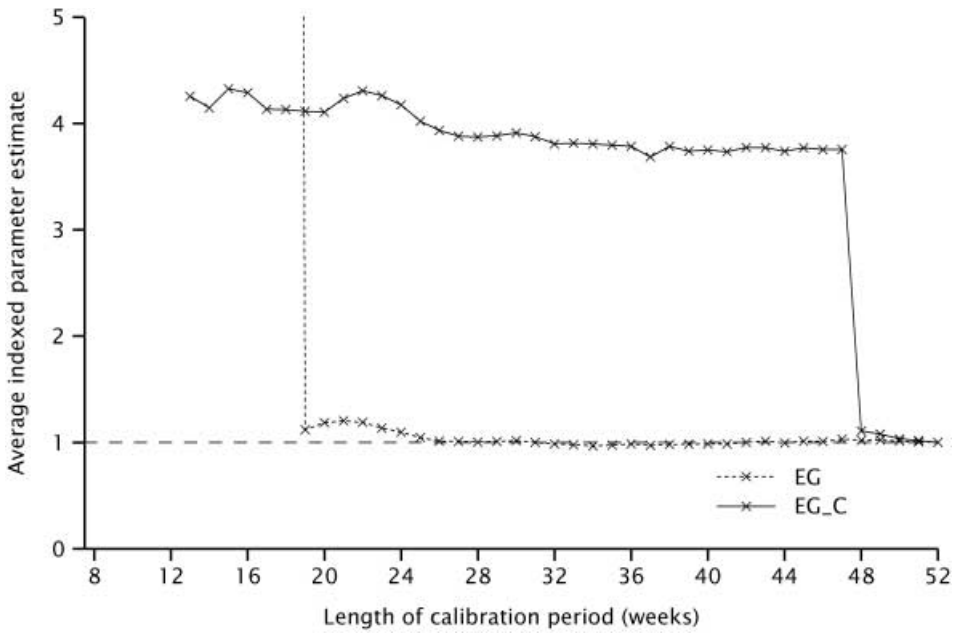

Figure 7. Parameter stability: gamma distribution variance parameter

and barely budge over all remaining calibration periods. (Thus we have further insight as to why the EG model generates poor forecasts when its model parameters are estimated using a short calibration period.) On the other hand, the EG_C model continues to overestimate the r parameter—albeit quite stable—until the calibration period reaches 48 weeks in length. This overestimation is not a particularly critical concern, since the α parameter adjusts to keep the timing distribution fairly accurate (as per Figure 5 and the forecasting results). In some cases, however, it may be desirable or important to ensure that all of the model parameters are maximally accurate and stable, in which case the pure EG model would be preferred.

The overall conclusion of the stability analysis is immediate: the only model specifications that pass the stability test for all parameters are the two exponential-gamma models. Just as we saw before, the EG_C model performs well for all calibration periods at least 12 weeks in length, while the simpler EG model is very poor until week 20 and very good beyond that point. It is encouraging to see such strong confirmation of the earlier forecasting results. Furthermore, the problems shown here for models involving the 'never triers' component provide clear evidence why we deem such models to be unreliable, despite the fact that their forecasts can be quite accurate in many cases. Finally it is good to see that the stability of the covariate parameters are reasonably invariant across the various model specifications. It is interesting that they are not adversely affected by the presence (or absence) of other model components. This finding demonstrates the value of performing this type of parameter stability analysis in conjunction with the focus on forecasting capabilities.

DISCUSSION AND CONCLUSIONS

The primary objectives of this research were (i) to study the impact on forecasting performance of incorporating marketing mix covariate effects into models of trial purchases, and (ii) to explore the trade-off between model calibration period and forecasting performance. In doing so, we have identified what can be considered the 'components' of a trial purchasing model with good forecasting properties. We summarize and synthesize our findings as a set of five principles that an analyst should be able to draw from our study. We then touch on a number of related issues that should be taken into consideration when generalizing beyond the scope of our study, and identify a set of future research questions.

(1) The underlying structural model, which dictates the time-to-trial for an individual panellist in each of our datasets, appears to be most consistent with a simple exponential process. On the surface, the data may not appear to be as clean and regular as a pure exponential process would imply, but this is due, in part, to the presence of heterogeneity and covariate effects (as well as random error).
(2) It is important to allow for consumer heterogeneity in the unobserved purchase rates. This observation might not have been immediately obvious from a quick first glance at Figure 1, but the subsequent analyses clearly showed that the two best models (in terms of forecast accuracy and parameter stability) feature a gamma mixing distribution for the exponential purchase rates.
(3) In contrast, the concept of a 'never triers' parameter seems reasonable at first glance, but does not hold up well under further scrutiny. On its own, this parameter acts as a weak proxy for a more comprehensive model of consumer heterogeneity, and when a separate heterogeneity distribution is included in the model, there appears to be a substantial confound between these two components. Based on forecasting accuracy, this parameter offers a negligible improvement (if

any) beyond the exponential-gamma models; moreover, when judged by parameter stability, the 'never triers' parameter fares very poorly. Its estimated values are highly unstable, even when long calibration periods are used.

(4) When marketing mix covariates such as advertising, coupons, and in-store promotions are available to the analyst, they can contribute significantly to the model's forecasting performance. This is especially true when the calibration period is fairly short. The performance of the exponential-gamma model with covariates is quite remarkable even with as few as 12 weeks of calibration data. The average MAPE for such a specification is roughly 10%, and it does not change dramatically even when 10–20 additional weeks of data are available to estimate the model parameters. Furthermore, the parameter stability analysis suggests that the estimated values of the covariate effects are fairly stable (especially with 20 or more weeks of calibration data), and relatively invariant across different model specifications. Thus, the covariates not only help explain some of the variation present in the time-to-purchase data, but they can be used to help guide policy decisions, such as the allocation of expenditures across markets and promotional vehicles.

(5) When covariates are unavailable (or untrustworthy), reliable forecasts can not be obtained until roughly 20 weeks of calibration data are available. However, after that point, the best no-covariate model (the exponential-gamma) becomes remarkably strong both in its forecasting accuracy and its parameter stability. On both criteria it actually surpasses the equivalent exponential-gamma model with covariates over most calibration periods beyond 20 weeks in length.

This last conclusion—the solid performance of the pure, no-covariate exponential-gamma model—is perhaps the most surprising finding in this paper. Although we found no problems at all with the inclusion of covariates in our various models, they apparently do not contribute much to a model's forecasting capabilities if the model is well-specified in the first place *and* a reasonable amount of data is available for parameter estimation. Of course, in many cases it is necessary to include covariates for diagnostic purposes, and they are absolutely essential if an analyst wishes to make forecasts before 20 weeks of data are available. But even when the analyst wishes to rely principally on a model with covariates, she should probably still run the pure EG model to get a quick and easy 'second opinion' about the forecast.

To elaborate on this latter point, the pure EG model is very easy to estimate using standard PC software (such as the Excel Solver). Furthermore, since its two parameters tend to show a high degree of stability, their estimated values could be databased to establish norms for future products or to serve as empirical priors for a Bayesian analysis of this forecasting problem. This could be an extremely useful exercise to help managers anticipate what market outcomes might be even before the new product is launched (e.g. Eskin and Malec, 1976). As Urban and Katz (1983) demonstrated for the ASSESSOR model, there are valuable benefits to be gained in the form of cumulative learning across multiple new product launches and their associated forecasting models.

It follows that a key area for future research involves the application of Bayesian methods to the problem of forecasting new product trial. In particular, Bayesian methods provide a framework for formally incorporating information about the market gained from previous new product launches. As such, they may greatly contribute to a model's forecasting performance, making it possible to generate sufficiently accurate forecasts of trial sales with a limited amount (i.e. less than 12 weeks) of in-market data.

Additional research questions can be asked about deeper issues embedded in the new product purchase process. One set of issues consists of subtleties involved in the trial process, for instance:

(1) why does the 'never triers' component fare so poorly in this case, and in what contexts might it be more helpful? (2) How well will the exponential-gamma models (as well as the others discussed here) capture heterogeneity across different geographic markets and/or different channels of distribution (e.g. grocery stores versus drug stores versus mass merchandisers)?; (3) How should the models be adapted to handle markets in which we observe distribution-build (i.e. markets that do not have the forced, 100% retail distribution that is present for all of the datasets used here)? There is a sparse amount of published research that covers these issues, especially in the CPG context, and a clear need for a better understanding of all of them.

Furthermore, there is a need to address issues that exist beyond the trial model, *per se*. We noted earlier that trial model tends to reflect the basic shape/nature of the subsequent repeat-purchase models, especially when repeat purchase behaviour is modelled using a series of 'depth-of-repeat' timing models (Eskin, 1973; Kalwani and Silk, 1980). Fader and Hardie (1999) have demonstrated that using the pure EG model in such a context results in a very robust model of repeat purchasing for a new CPG product (in spite of the theoretical concerns previously raised in footnote 2). A natural extension to this work would be to replace the core EG model with the EG_C model to arrive at a model of repeat purchasing for a new CPG product that incorporates the effects of marketing mix covariates. The research objectives would be to examine what impact these covariates have on the accuracy of the repeat sales forecasts, and to explore whether the inclusion of marketing mix covariates enables us to reduce the length of the model calibration period (i.e. shorten the test market). In any event, the specific form and implementation of the repeat purchase model is outside the scope of this paper, but it is encouraging to know that the 'winning' trial model discussed here lends itself to a variety of potentially useful repeat models.

APPENDIX A: INCORPORATING THE EFFECTS OF COVARIATES

Over the past 30 years, researchers in a number of disciplines such as biostatistics, economics, and sociology have developed methods for studying event-time data (e.g. Cox and Oakes, 1984; Kalbfleisch and Prentice, 1980; Lancaster, 1990; Lawless, 1982). At the heart of these methods is the hazard rate function,

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{A1}$$

which specifies the instantaneous rate of the event (e.g. trial) occurring at $T = t$ conditional on it not having occurred up to time t. The hazard rate function and cdf are mathematically equivalent ways of specifying the distribution of a continuous nonnegative random variable. Because $F(0) = 0$, it follows from (A1) that

$$\begin{aligned} F(t) &= 1 - \exp\left(1 - \int_0^t h(u)du\right) \\ &= 1 - \exp(-H(t)) \end{aligned} \tag{A2}$$

where $H(t)$ is called the integrated hazard function.

A popular, easily interpretable method for incorporating the effects of exogenous covariates into event-time models is the proportional hazards approach. In this framework, the covariates have a multiplicative effect on the hazard rate. More specifically, let $F_0(t|\theta)$ be the so-called 'baseline' cdf

for the distribution of an individual's time-to-trial, and $f_0(t|\theta)$ and $h_0(t|\theta)$ the associated pdf and hazard rate function. The most common formulation of the proportional hazards specification states that

$$h(t\,|\,\theta,\, \boldsymbol{x}(t),\, \beta) = h_0(t\,|\,\theta)\exp[\boldsymbol{\beta}'\mathbf{x}(t)]$$

where $\boldsymbol{x}(t)$ denotes the vector of covariates at time t and $\boldsymbol{\beta}$ denotes the effects of these covariates. It follows from (A2) that the with-covariates cdf for the distribution of time-to-trial is given by

$$\begin{aligned} F(t\,|\,\theta,\, \mathbf{X}(t),\, \beta) &= 1 - \exp\left(-\int_0^t h(u\,|\,\theta,\, \boldsymbol{x}(t),\, \boldsymbol{\beta})du\right) \\ &= 1 - \exp(-H(t\,|\,\theta,\, \boldsymbol{X}(t),\, \boldsymbol{\beta})) \end{aligned}$$

where $\mathbf{X}(t)$ represents the covariate path up to time t, i.e. $\{\boldsymbol{x}(u)\colon 0 < u \le t\}$.

Assuming the time-varying covariates remain constant *within* each unit of time (e.g. week),

$$\begin{aligned} H(t|\theta, \boldsymbol{X}(t), \boldsymbol{\beta}) &= \int_0^1 h(u)du + \int_1^2 h(u)du + \ldots + \int_{\mathrm{Int}(t)}^t h(u)du \\ &= \exp[\boldsymbol{\beta}'\boldsymbol{x}(1)]\int_0^1 h_0(u)du + \exp[\boldsymbol{\beta}'\boldsymbol{x}(2)]\int_1^2 h_0(u)du + \ldots \\ &\quad + \exp[\boldsymbol{\beta}'\boldsymbol{x}(\mathrm{Int}(t+1))]\int_{\mathrm{Int}(t)}^t h_0(u)du \\ &= \sum_{i=1}^{\mathrm{Int}(t)} [\ln[1 - F_0(i-1|\theta)] - \ln[1 - F_0(i|\theta)]]\exp[\boldsymbol{\beta}'\boldsymbol{x}(i)] \\ &\quad + [\ln[1 - F_0(\mathrm{Int}(t)|\theta)] - \ln(1 - F_0(t|\theta))]\exp[\boldsymbol{\beta}'\boldsymbol{x}(\mathrm{Int}(t+1))] \end{aligned} \tag{A3}$$

since $\int_{i-1}^{i} h_0(u)du = -\ln[1 - F_0(u)]_{i-1}^{i} = \ln[1 - F_0(i-1)] - \ln[1 - F_0(i)]$.

For specific baseline distributions, the above expression can be simplified; for example, if $F_0(t|\theta)$ is exponential with rate parameter θ, we have

$$\begin{aligned} H(t\,|\,\theta,\, X(t),\, \boldsymbol{\beta}) &= \theta\left\{\sum_{i=1}^{\mathrm{Int}(t)} \exp[\boldsymbol{\beta}'\boldsymbol{x}(i)] + [t - \mathrm{Int}(t)]\exp[\boldsymbol{\beta}'\boldsymbol{x}(\mathrm{Int}(t+1))]\right\} \\ &\equiv \theta A(t) \end{aligned}$$

Therefore the cdf of the with-covariates extension of the exponential distribution is

$$F(t\,|\,\theta,\, \boldsymbol{X}(t),\, \boldsymbol{\beta}) = 1 - \exp(-\theta A(t)) \tag{A4}$$

When $\boldsymbol{\beta} = \mathbf{0}$ (i.e. the covariates are omitted), $A(t) = t$ and (A4) reduces to the cdf of the exponential distribution (i.e. the baseline cdf).

(*Note*: Appendix B is available at http://brucehardie.com/papers/fhz_appendix_b.pdf)

ACKNOWLEDGEMENTS

The authors thank Information Resources, Inc. for its assistance and the provision of data, and J. Scott Armstrong for his comments on an earlier version of the paper. The second author acknowledges the support of the London Business School Research & Materials Development Fund and the LBS Centre for Marketing.

REFERENCES

Anscombe FJ. 1961. Estimating a mixed-exponential response law. *Journal of the American Statistical Association* **56**: 493–502.

Armstrong JS, Collopy F. 1992. Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting* **8**: 69–80.

Baum J, Dennis KER. 1961. The estimation of the expected brand share of a new product. *VIIth ESOMAR/WAPOR Congress*.

Bass FM, Jain D, Krishnan T. 2000. Modeling the marketing-mix influence in new-product diffusion. In *New-Product Diffusion Models*, Mahajan V, Muller E, Wind Y (eds). Kluwer Academic Publishers: Boston, MA.

Bass FM, Krishnan TV, Jain DC. 1994. Why the Bass model fits without decision variables. *Marketing Science* **13**: 203–223.

Bottomley PA, Fildes R. 1998. The role of prices in models of innovation diffusion. *Journal of Forecasting* **17**: 539–555.

Broadbent S. 1984. Modelling with Adstock. *Journal of the Market Research Society* **26**: 295–312.

Cox DR, Oakes D. 1984. *Analysis of Survival Data*. Chapman & Hall: London.

Curry DJ. 1993. *The New Marketing Research Systems*. John Wiley: New York.

Eskin GJ. 1973. Dynamic forecasts of new product demand using a depth of repeat model. *Journal of Marketing Research* **10**: 115–129.

Eskin GJ. 1974. Causal structures in dynamic trial-repeat forecasting models. *1974 Combined Proceedings, Series No. 36*, American Marketing Association: Chicago, IL.

Eskin GJ, Malec J. 1976. A model for estimating sales potential prior to the test market. *Proceeding 1976 Fall Educators' Conference, Series No. 39*, American Marketing Association: Chicago, IL.

Fader PS, Hardie BGS. 1999. Investigating the properties of the Eskin/Kalwani & Silk model of repeat buying for new products. In *Marketing and Competition in the Information Age*, Proceedings of the 28th EMAC Conference, 11–14 May, Hildebrandt L, Annacker D, Klapper D (eds). Humboldt University: Berlin.

Fildes R. 1992. The evaluation of extrapolative forecasting methods. *International Journal of Forecasting* **8**: 81–98.

Fourt LA, Woodlock JW. 1960. Early prediction of market success for new grocery products. *Journal of Marketing* **25**: 31–38.

Gupta S, Morrison DG. 1991. Estimating heterogeneity in consumers' purchase rates. *Marketing Science* **10**: 264–269.

Hardie BGS, Fader PS, Wisniewski M. 1998. An empirical comparison of new product trial forecasting models. *Journal of Forecasting* **17**: 209–229.

Herniter J. 1971. A probabilistic market model of purchase timing and brand selection. *Management Science* **18**: P102–P113.

Kalbfleisch JD, Prentice RL. 1980. *The Statistical Analysis of Failure Time Data*. John Wiley: New York.

Kalwani M, Silk AJ. 1980. Structure of repeat buying for new packaged goods. *Journal of Marketing Research* **17**: 316–322.

Lancaster T. 1990. *The Econometric Analysis of Transition Data*. Cambridge University Press: Cambridge.

Lawless JF. 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley: New York.

Makridakis S. 1993. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting* **9**: 527–529.

Massy WF. 1969. Forecasting the demand for new convenience products. *Journal of Marketing Research* **6**: 405–412.

Morrison DG, Schmittlein DC. 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.
Nakanishi M. 1973. Advertising and promotion effects on consumer response to new products. *Journal of Marketing Research* **10**: 242–249.
Urban GL, Katz GM. 1983. Pre-test market models: validation and managerial implications. *Journal of Marketing Research* **20**: 221–234.

Authors' biographies:

Peter S. Fader is a Professor of Marketing at the Wharton School of the University of Pennsylvania. His research focuses on developing forecasting models for a variety of domains, including new product performance, Internet shopping/buying activities, movements of celestial bodies, seasonal leaf-colouration patterns, and children's shoe sizes.

Bruce G. S. Hardie is an Assistant Professor of Marketing at London Business School. His primary research interest is in the area of stochastic modelling in marketing. Current projects include the development of probability models for new product sales forecasting and understanding online consumer behaviour.

Robert Zeithammer is a PhD student at the Sloan School of Management at MIT. His research focuses on all things new including new product sales, new Bayesian statistical methodologies, and new auction-driven markets facilitated by the Internet.

Authors' addresses:

Peter S. Fader, Marketing Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, USA.

Bruce G. S. Hardie, London Business School, Regent's Park, London NW1 4SA, UK.

Robert Zeithammer, E56-345e, MIT Sloan School, Cambridge, MA 02139, USA.

