Forecasting New Product Trial in a Controlled Test Market Environment

Peter S. Fader
Bruce G. S. Hardie
Robert Zeithammer [1]

April 2002

[1]Peter S. Fader is Associate Professor of Marketing at the Wharton School, University of Pennsylvania, Bruce G. S. Hardie is Assistant Professor of Marketing, London Business School, and Robert Zeithammer is a PhD candidate at MIT. The order of authorship is alphabetical. 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.

Abstract

Central to the development of many new consumer packaged goods is the use of an in-market test (e.g., a controlled test market) prior to launch. A number of market researchers have proposed forecasting models that generate a number of useful diagnostics about the product's likely performance and also enable managers to shorten the length of these test markets. 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 test market).

Rather than "tack-on" covariate effects to existing models of new product trial, we develop from first principles a set of models that enable us to systematically explore the impact of three model "components" on forecasting performance: 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 modeled, and iii) whether or not the effects of marketing decision variables are incorporated. Furthermore, we systematically explore the impact of the length of the test market on forecasting performance.

Having identified that the distribution of time-to-trial for an individual is best characterized by the exponential distribution, we find that it is critically important to capture heterogeneity (via a gamma distribution across households), and that the inclusion of covariate effects is often a useful addition, especially for models calibrated on fewer than 20 weeks of data. The "never triers" parameter proves to be largely ineffective, mostly because of its apparent redundancy with the heterogeneity distribution. We provide detailed evidence for these conclusions, and link them to further research issues.

1 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 reduce the duration of a market test reduces the cost of the test itself as well as the opportunity costs of not going national earlier (assuming the test would result in a "go national" decision).

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 modeling 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., 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

[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, Fader, and Wisniewski 1998).

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), 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 modeling 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, Jain, and Krishnan 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 the model calibration period and the

[2] A problem with such this depth-of-repeat approach is that it can result in misleading inferences about buyer behavior, 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 modeling framework for a number of practical reasons, principally the quality of its forecasts (Fader and Hardie 1999).

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.

2 Model Development

When conducting a test market, it is standard practice to monitor the performance of the new product using panel data. For each household i (in a panel of N households), we observe t_i, the time at which it made a trial purchase. (The zero point of the time scale corresponds to the time of the introduction of the new product.) The empirical market penetration curve for the new product is

$$\hat{F}(t) = N^{-1} \left(\# \text{ of } t_i \leq t\right) ,$$

which can be interpreted as the aggregate-level empirical cumulative distribution function of trial purchase times.

In building a trial purchase model, we specify a parametric form for $F(t)$. 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 modeled 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 modeling approach in which they make a set of assumptions about consumer behavior, 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:

i. 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*.)

ii. When specifying the structural model, we take the perspective of a single panelist. As we move from one panelist to all panelists 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.

iii. 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 a 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 panelists who do not have children (or grandchildren) under 3 years old. This implies that the assumed cdf for a panelist's time-to-trial only applies to those panelists that will eventually trial 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 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 modeling 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 in event-time models is the proportional hazard approach. (See

Appendix A for a brief review.) This leads to $F(t|\theta, \boldsymbol{X}(t), \boldsymbol{\beta})$, an individual-level with-covariate cdf for the distribution of time-to-trial, where $\boldsymbol{X}(t)$ 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

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

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 $F(t|\theta, \boldsymbol{X}(t), \boldsymbol{\beta})$, $g(\theta)$, 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, 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 $F(t|\theta, \boldsymbol{X}(t), \boldsymbol{\beta}) = 1 - \exp\big(-\theta A(t)\big)$ 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 modeled, 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.,

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

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 1. 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 1 about here]

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.

3 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 eight markets, geographically dispersed across the U.S.; six of these are targetable TV markets (Pittsfield, MA, Marion, IN, Eau Claire, WI, Midland, TX, Grand Junction, CO, and Cedar Rapids, IA), the other two are non-targetable TV markets (Visalia, CA and Rome, GA). (See Curry (1993) for further details of the *BehaviorScan* service.) We have five datasets (labeled 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 panelist 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 panelists 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

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

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 1.) 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 $\times$ 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 8–51 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 $\times$ calibration period $\times$ dataset $(8 \times 44 \times 5)$ combinations. These will then be analyzed 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 $\times$ calibration period $\times$ 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.)

4 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 parameters. 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 (forecasting ability in section 4.1 and parameter variation in section 4.2), 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 behavior.

4.1 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 un-adorned 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 roughly 55%.

[Figure 1 about here]

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 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 1). 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 about here]

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 favor 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.

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.

[Figure 3 about here]

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 modeled, and the forthcoming parameter stability analysis (section 4.2) will clearly show why we favor the model with one component over both.

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.

[Figure 4 about here]

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.

4.2 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 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 analyze parameter variation across different model specifications and calibration periods, we created indexed parameter estimates by dividing all parameters for each of the 1760 model $\times$ calibration period $\times$ 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.

[Figure 5 about here]

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, regardless 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 section 4.1 — 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 behavior 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.

[Figure 6 about here]

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).

[Figure 7 about here]

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

[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.

these two models. The estimated values of r for the EG model move almost precisely to the 52-week values, 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.

5 Discussion and Conclusions

The primary objectives of this research were i) to study the impact on forecasting performance of incorporating marketing mix covariate effects in 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.

i. The underlying structural model, which dictates the time-to-trial for an individual panelist 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).

ii. 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.

iii. 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.

iv. 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.

v. 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 vs. drug stores vs. mass merchandisers)?; and (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 behavior is modeled 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. 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 occured 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

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

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

A popular, easily interpretable method for incorporating the effects of exogenous covariates in 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

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

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

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

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

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

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

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

$$H(t|\theta, X(t), \boldsymbol{\beta}) = \theta\Big\{\sum_{i=1}^{\text{Int}(t)} \exp[\boldsymbol{\beta}'\boldsymbol{x}(i)] + [t - \text{Int}(t)] \exp[\boldsymbol{\beta}'\boldsymbol{x}(\text{Int}(t+1))]\Big\}$$
$$= \theta A(t)$$

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

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

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).

Appendix B: Identifying the Structural Model

At the heart of any model of trial purchasing lies the specification of the structural model; that is, the distribution of an individual's time-to-trial. As previously noted, developers of the various published trial forecasting models have all assumed a particular structural model without providing any justification for their choice. As we seek to gain a thorough understanding of trial forecasting models, it is important that we build on a solid foundation, i.e., start with the "correct", rather than assumed, structural model, $F_0(t|\theta)$. In this appendix we report on a careful empirical investigation designed to identify the correct form of the structural model for models of time-to-first purchase for consumer packaged goods.[4] Within the marketing literature on purchase timing, a number of researchers have explored the issue of what distribution should be used to characterize interpurchase times (e.g., Gupta (1991), Jain and Vilcassim (1991)). The findings of this stream of work are mixed in terms of determining what distribution should be used to model interpurchase times. Furthermore, this work has focused on established CPG products. It is therefore important that we perform our own analysis in this new CPG product setting.

It is well known that a failure to control for unobserved heterogeneity in the empirical investigation of a structural model results in biases in the estimated hazard rate function; in particular, there is bias towards negative duration dependence (Heckman and Singer 1984a). For example, in fitting a Weibull distribution to some data, we may find that the estimated value c is less than 1.0 and conclude that there is a decreasing hazard rate; in actual fact, the observed data could be the realization of a heterogeneous exponential process, and our failure to account for heterogeneity would lead us to draw incorrect conclusions. (See Vaupel and Yashin (1985) for a complete discussion of other incorrect conclusions that can be drawn from fitting homogeneous models to heterogeneous data.)

It is therefore standard practice to use some parametric distribution, $g(\theta)$, to control for unobservables. (The choice of $g(\theta)$ is typically justified on the grounds of mathematical convenience.) What is less well-known is that inferences about the structural model are

[4] In deeming a particular cdf, $F_0(t|\theta)$, to be the "correct" structural model, we are conditioning on a *given set of candidate structural models*. If the true structural model is not one of the candidate models, we would not be able to identify it. However, the set of candidate models considered in this paper represent the set of models considered in the marketing literature and can therefore be viewed as sufficiently exhaustive.

sensitive to assumptions made about the distribution of unobserved heterogeneity. While $F_0(t|\theta)$ and $g(\theta)$ are unobservable, we can estimate the mixture $F(t)$. Unfortunately it not possible to uniquely identify $F_0(t|\theta)$ and $g(\theta)$, given $F(t)$; for any $F(t)$, one can produce different structural models that, along with a matching $g(\theta)$, result in the same $F(t)$ — see, for example, Heckman and Singer (1984b, p. 274).

A good marketing-related example of this problem is found by examining the so-called "Bass model", where the hazard rate of the distribution of adoption times is of the form $p + qF(t)$. Note that there is no explicit accounting for heterogeneity and therefore the model should be interpreted as describing a population of homogeneous buyers. However, as shown by Bemmaor (1994), it can be derived/interpreted as the mixture distribution associated with a shifted Gompertz structural model (i.e., an individual consumer's time to adoption follows a shifted Gompertz distribution) and an exponential mixing distribution to account for consumer heterogeneity. Therefore the "Bass model" may fit observed adoption time data reasonably well, but it is not possible to distinguish between these alternative explanations.

With this in mind, we cannot say that Hardie et al.'s (1998) support for EG model implies that time-to-trial at the individual-level follows the exponential distribution (with heterogeneity in trial rates captured by a gamma distribution). It could be the case that we have a homogeneous group of buyers, each of whose time-to-trial follows the Lomax or "Pareto of the second kind" distribution (which are other names by which the EG model goes by in the statistics literature).

Heckman and Singer (1984b) propose the use of a non-parametric approach to controlling for the unobserved heterogeneity, which allows us to investigate the correct form of the structural model without imposing arbitrary parametric specifications for $g(\theta)$ that could bias the identification of $F_0(t|\theta)$. In particular, they propose that $g(\theta)$ be specified as a discrete pdf with K support points at locations θ_k and associated probability mass π_k $(k = 1, \ldots, K, \sum_k \pi_k = 1)$.[5]

In our empirical analysis, we will utilize this approach to control for unobserved heterogeneity as we attempt to identify the "correct" structural model for the trial process. Our

[5] In this study, we assume $g(\theta)$ is univariate; in principle, it could be multivariate.

search for the correct model proceeds as follows: we identify a set of candidate structural models, and fit them to a number of datasets using the Heckman and Singer approach to control for unobserved heterogeneity. Using certain fit criteria, we identify which structural model best fits the observed trial pattern for each dataset. From this we identify the "best" overall structural model.

The candidate structural models are identified from the existing literature. The simplest model is the exponential distribution, as used by Anscombe (1961) and indirectly by Fourt and Woodlock (1960) and Eskin (1973, 1974). Such a distribution is consistent with the Poisson counting process that lies at the heart of the familiar NBD model (Ehrenberg 1959; Morrison and Schmittlein 1988), which is used to characterize the distribution of repeat buying for established CPG products. A generalization of this distribution is the Erlang-2 distribution, as proposed by Greene (1982) and used by Herniter (1971) in his empirical analysis. In an established product setting, this distribution has been used by Gupta (1991), and is the timing model that corresponds to the counting process associated with the CNBD model (Chatfield and Goodhardt 1973, Schmittlein and Morrison 1983). Another generalization of the exponential distribution is the Weibull model, as used by Massy (Massy 1968, 1969; Massy, Montgomery, and Morrison 1970). Finally, we consider the Erlang-k distribution, as proposed by Herniter (1971). Rather than directly using the Erlang-k distribution, we use the gamma distribution, which relaxes the restriction on the Erlang-k distribution that the shape parameter be a positive integer k. The pdfs associated with this set of structural models are presented in Table B1.

[Table B1 about here]

We perform this analysis on the five *BehaviorScan* test market datasets described in Section 3. Given the interval-censored nature of these data, the likelihood functional, conditional on θ_k, is:

$$
\begin{aligned}
L(\theta_k, \boldsymbol{\beta}|\text{data}) &= \left[1 - F(52|\theta_k, \boldsymbol{\beta})\right]^{\left(N - \sum_{i=1}^{52} n_i\right)} \prod_{i=1}^{52} \left[F(i|\theta_k, \boldsymbol{\beta}) - F(i-1|\theta_k, \boldsymbol{\beta})\right]^{n_i} \\
&= \exp\left[-H(52|\theta_k, \boldsymbol{X}(52), \boldsymbol{\beta})\right]^{\left(N - \sum_{i=1}^{52} n_i\right)} \\
&\quad \times \prod_{i=1}^{52} \left\{\exp[-H(i-1|\theta_k, \boldsymbol{X}(i-1), \boldsymbol{\beta})] - \exp[-H(i|\theta_k, \boldsymbol{X}(i), \boldsymbol{\beta})]\right\}^{n_i}
\end{aligned}
$$

where n_i is the number of triers in week i (i.e., # of $t_i \in (i-1, i]$), N is the number of households in the panel, and

$$
H(t|\theta_k, X(i), \beta) = \sum_{i=1}^{t} \left[\ \ln\left[1 - F_0(i-1|\theta_k)\right] - \ln\left[1 - F_0(i|\theta_k)\right]\ \right] \exp[\boldsymbol{\beta}'\boldsymbol{x}(i)]
$$

For K support points, the overall likelihood function is

$$
L = \sum_{k=1}^{K} \pi_k L(\theta_k, \beta)
$$

Note that this formulation of the likelihood function does not force a "never triers" term (i.e., p); for the class of structural models considered here, this would correspond to a mass point of size $1 - p$ located at $\theta = 0$. Empirically, the data may "request" a mass point at zero. However, at this stage of the analysis, where we wish to identify the "correct" structural model, it would be counter-productive to force a mass point at zero as this would represent a constraint on the distribution of unobserved heterogeneity.

For each dataset, we fit the time-to-first purchase model $F(t)$ associated with each of the four structural models under consideration. For each structural model, the exact specification of the non-parametric $g(\theta)$ is determined by using the heuristic of choosing the number of mass points to minimize a log-likelihood-based fit measure. The particular measure used is Bozdogan's (1987) CAIC measure: CAIC $= -2LL + q[\ln(N) + 1]$, where LL = log-likelihood, q = number of model parameters, and N = number of observations (i.e., the panel size).

Specifications with $K = 1$, 2, and 3 support points were estimated for each dataset $\times$ structural model combination. The summary results are reported in Table B2; in every

case, CAIC was minimized with one or two support points, so the numbers corresponding to 3 support points are not reported.

[Table B2 about here]

Before interpreting these results, it is important to note that, for some of the datasets, we were unable to obtain interior solutions in the parameter space for the model specifications associated with Weibull and gamma structural models. Despite repeated attempts to re-estimate the models, the same corner solution was obtained. In all cases, this was associated with one or more of the θ_k tending to zero (i.e., a "never triers" component).

The logic followed in identifying the "correct" structural model is to compare the fit (in terms of CAIC) of the various structural models for each dataset. Within each structural model, we consider the non-parametric heterogeneity specification (be it one or two support points) that best fits the data; i.e., minimizes CAIC. In all cases, the exponential structural model best fits the data, and we can therefore conclude that it is the "correct" structural model to use when developing models of time-to-first purchase, at least for the range/type of datasets considered here.

Aside from this main result, there are two interesting points to note. First, in all cases, the estimate of the shape parameter of the gamma structural model is less than or equal to 1.0 (i.e., a monotone decreasing or constant hazard rate); this suggests there is no support for the Erlang-k model as proposed by Herniter (1971) and Greene (1982). Second, we consider the issue of how to specify $g(\theta)$ in our subsequent model building work. It is important to note that Heckman and Singer's (1984b) proposal to use a non-parametric $g(\theta)$ was strictly for the purposes of identifying the structural model; once the structural model has been identified, there is nothing stopping the modeler from using a flexible parametric mixing distribution for subsequent work. We note from the above analysis that the non-parametric specification of $g(\theta)$ requires very few support points. This suggests that, for the data associated with new consumer packaged goods products, we can use a simple unimodal parametric distribution to capture unobserved heterogeneity. This provides support for our use of a gamma mixing distribution when we develop models

of trial purchasing.

Readers familiar with the hazard rate modeling literature may be wondering we haven't used models with a nonparametric baseline hazard function and parametric unobserved heterogeneity (e.g., Han and Hausman (1990), Meyer (1990), Vanhuele et al. (1995)). While such an approach is very useful when there is a need to make inferences about the shape of the hazard rate function for the purpose of hypothesis testing, it becomes useless in a forecasting setting. Should there be evidence of duration dependence, the need to make predictions well beyond the calibration period implicitly requires us to know the exact form the hazard rate function well into the future. This can be very difficult with a nonparametric baseline hazard function, whereas it is automatic with a parametric baseline hazard function. Given that forecasting is central to this paper, we have focused on parametric baseline hazard functions.

References

Anscombe, F. J. (1961), "Estimating a Mixed-Exponential Response Law," *Journal of the American Statistical Association*, **56** (September), 493–502.

Armstrong, J. Scott and Fred Collopy (1992), "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons," *International Journal of Forecasting*, **8** (1), 69–80.

Baum, J. and K. E. R. Dennis (1961), "The Estimation of the Expected Brand Share of a New Product," *VIIth ESOMAR/WAPOR Congress*.

Bass, Frank M., Dipak Jain, and Trichy Krishnan (2000), "Modeling the Marketing-Mix Influence in New-Product Diffusion," in Vijay Mahajan, Eitan Muller, and Yoram Wind (eds.), *New-Product Diffusion Models*, Boston: Kluwer Academic Publishers, 99–122.

Bemmaor, Albert C. (1994), "Modeling the Diffusion of New Durable Goods: Word-of-Mouth Effect Versus Consumer Heterogeneity," in Gilles Laurent, Gary L. Lilien, and Bernard Pras (eds.), *Research Traditions in Marketing*, Boston: Kluwer Academic Publishers, 201–223.

Bottomley, Paul A. and Robert Fildes (1998), "The Role of Prices in Models of Innovation Diffusion," *Journal of Forecasting*, **17** (December), 539–555.

Bozdogan, Hamparsum (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions," *Psychometrika*, **52** (September), 345–370.

Broadbent, Simon (1984), "Modelling with Adstock," *Journal of the Market Research Society*, **26** (4), 295–312.

Chatfield, C. And G. J. Goodhardt (1973), "A Consumer Purchasing Model with Erlang Inter-Purchase Times," *Journal of the American Statistical Association*, **68** (December), 828–835.

Cox, D.R. and D. Oakes (1984), *Analysis of Survival Data*, London: Chapman & Hall.

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

Ehrenberg, A. S. C. (1959), "The Pattern of Consumer Purchases," *Applied Statistics*, **8** (March), 26–41.

Eskin, Gerald J. (1973), "Dynamic Forecasts of New Product Demand Using a Depth of Repeat Model," *Journal of Marketing Research*, **10** (May), 115–129.

Eskin, Gerald J. (1974), "Causal Structures in Dynamic Trial-Repeat Forecasting Models," *1974 Combined Proceedings, Series No. 36*, Chicago, IL: American Marketing Association, 198–201.

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, 230–233.

Fader, Peter S. and Bruce G. S. Hardie (1999), "Investigating the Properties of the Es-

kin/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.

Fildes, Robert (1992), "The Evaluation of Extrapolative Forecasting Methods," *International Journal of Forecasting*, **8** (1), 81–98.

Fourt, Louis A. and Joseph W. Woodlock (1960), "Early Prediction of Market Success for New Grocery Products," *Journal of Marketing*, **25** (October), 31–38.

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

Gupta, Sunil (1991), "Stochastic Models of Interpurchase Time with Time-Dependent Covariates," *Journal of Marketing Research*, **28** (February), 1–15.

Gupta, Sunil and Donald G. Morrison (1991), "Estimating Heterogeneity in Consumers' Purchase Rates," *Marketing Science*, **10** (Summer), 264–269.

Han, Aaron and Jerry A. Hausman (1990), "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*, **5** (January–March), 1–28.

Hardie, Bruce G. S., Peter S. Fader, and Michael Wisniewski (1998), "An Empirical Comparison of New Product Trial Forecasting Model," *Journal of Forecasting*, **17** (June–July), 209–229.

Heckman, James J. and Burton Singer (1984a), "Econometric Duration Analysis," *Journal of Econometrics*, **24** (January/February), 63–132.

Heckman, J. and B. Singer (1984b), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, **52** (March), 271–320.

Herniter, Jerome (1971), "A Probabilistic Market Model of Purchase Timing and Brand Selection," *Management Science*, **18** Part II (December), P102–P113.

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

Kalbfleisch, John D. and Ross L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons, Inc.

Kalwani, Manohar and Alvin J. Silk (1980), "Structure of Repeat Buying for New Packaged Goods," *Journal of Marketing Research*, **17** (August), 316–322.

Lancaster, Tony (1990), *The Econometric Analysis of Transition Data*, Cambridge: Cambridge University Press.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons, Inc.

Makridakis, Spyros (1993), "Accuracy Measures: Theoretical and Practical Concerns," *International Journal of Forecasting*, **9** (4), 527–529.

Massy, William F. (1968), "Stochastic Models for Monitoring New-Product Introduction," in Frank M. Bass, Charles W. King, and Edgar A. Pessemier (eds.), *Applications of the Sciences in Marketing Management*, New York: John Wiley and Sons, 85–111.

Massy, William F. (1969), "Forecasting the Demand for New Convenience Products," *Journal of Marketing Research*, **6** (November), 405–412.

Massy, William F., David B. Montgomery, and Donald G. Morrison (1970), *Stochastic Models of Buying Behavior*, Cambridge, MA: The MIT Press.

Meyer, Bruce D. (1990), "Unemployment Insurance and Unemployment Spells," *Econometrica*, **58** (July), 757–782.

Morrison, Donald G. and David C. Schmittlein (1988), "Generalizing the NBD Model for Customer Purchases: What Are the Implications and Is It Worth the Effort?" *Journal of Business and Economic Statistics*, **6** (April), 145–159.

Nakanishi, Masao (1973), "Advertising and Promotion Effects on Consumer Response to New Products," *Journal of Marketing Research*, **10** (August), 242–249.

Schmittlein, David C. and Donald G. Morrison (1983), "Prediction of Future Random Events With the Condensed Negative Binomial Distribution," *Journal of the American Statistical Association*, **78** (June), 449–456.

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

Vanhuele, Marc, Marnik G. Dekimpe, Sunil Sharma, and Donald G. Morrison (1995), "Probability Models for Duration: The Data Don't Tell the Whole Story," *Organizational Behavior and Human Decision Processes*, **62** (April), 1–13.

Vaupel, James W. and Anatoli I. Yashin (1985), "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics," *The American Statistician*, 39 (August), 176–185.

Table 1: 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

Table B1: Alternative Structural Models

Structural Model	Probability Density Function
exponential	$f_0(t\|\theta) = \theta e^{-\theta t}$
Erlang-2	$f_0(t\|\theta) = \theta^2 t e^{-\theta t}$
Weibull	$f_0(t\|\theta, c) = c\theta^c t^{c-1} e^{-(\theta t)^c}$
gamma	$f_0(t\|\theta, \mu) = \dfrac{\theta^\mu t^{\mu-1} e^{-\theta t}}{\Gamma(\mu)}$

Table B2: Model Fit (CAIC) for the Various Specifications

Structural Model	# Support Points	# Model Params	Data Set				
			A (N=566)	B (N=1300)	C (N=721)	D (N=2273)	E (N=2946)
exponential	1	4	2046.8	2570.3	1832.1	2333.0	7822.5
	2	6	2025.3	2551.2	1801.9	2349.2	7832.9
Erlang-2	1	4	2192.9	2879.5	2015.7	2404.9	8065.3
	2	6	**	2664.2	1988.2	2361.7	7856.9
Weibull	1	5	2026.3	**	**	2340.7	7827.8
	2	7	2028.1	**	**	**	7839.5
gamma	1	5	2026.9	**	**	2340.7	7828.2
	2	7	**	**	**	2358.2	**

** = corner solution

Figure 1:
Forecasting errors: All models (average across 5 datasets)

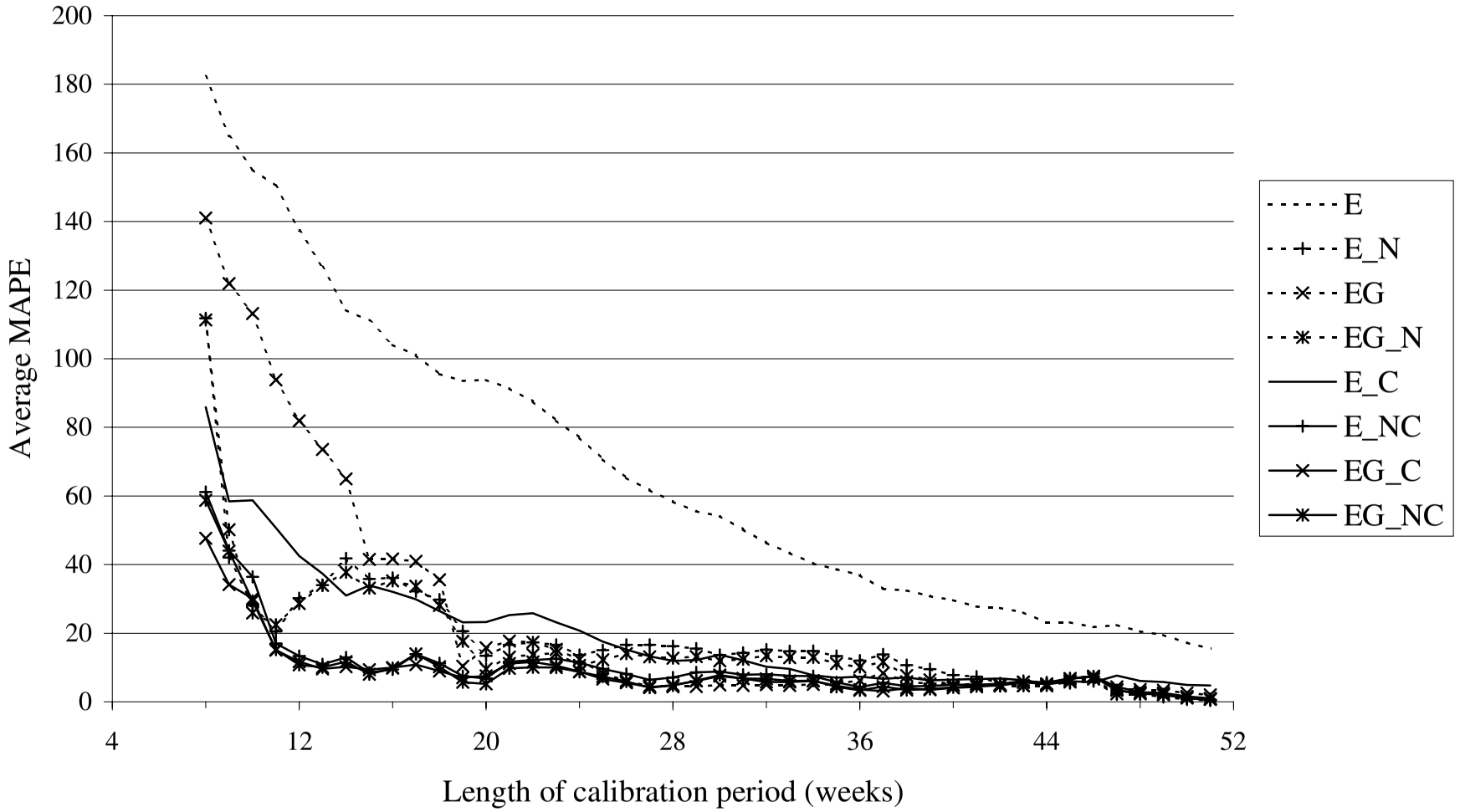

Figure 2:
Forecasting errors: Models with covariates only

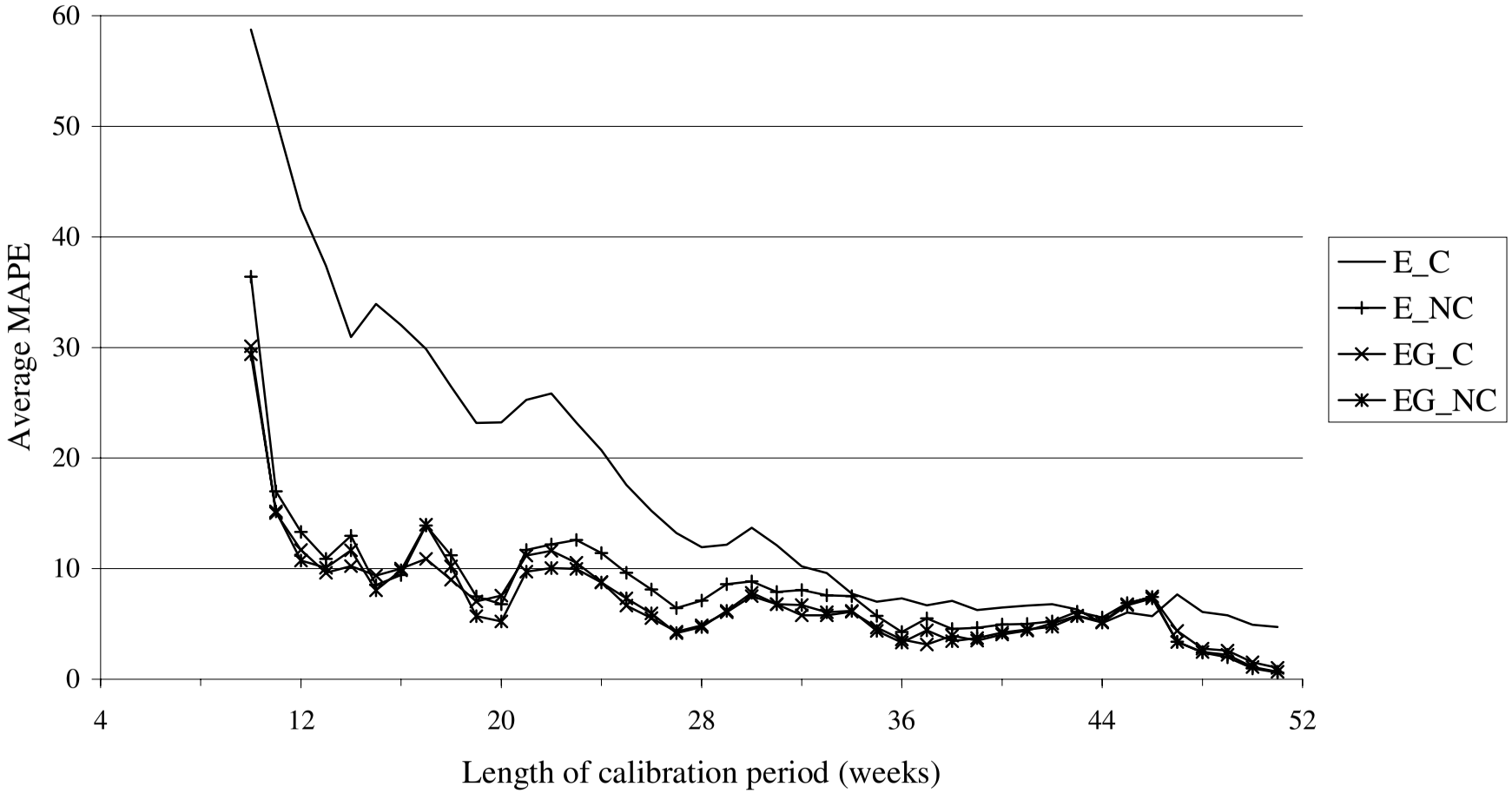

Figure 3:
Forecasting errors: Models without covariates

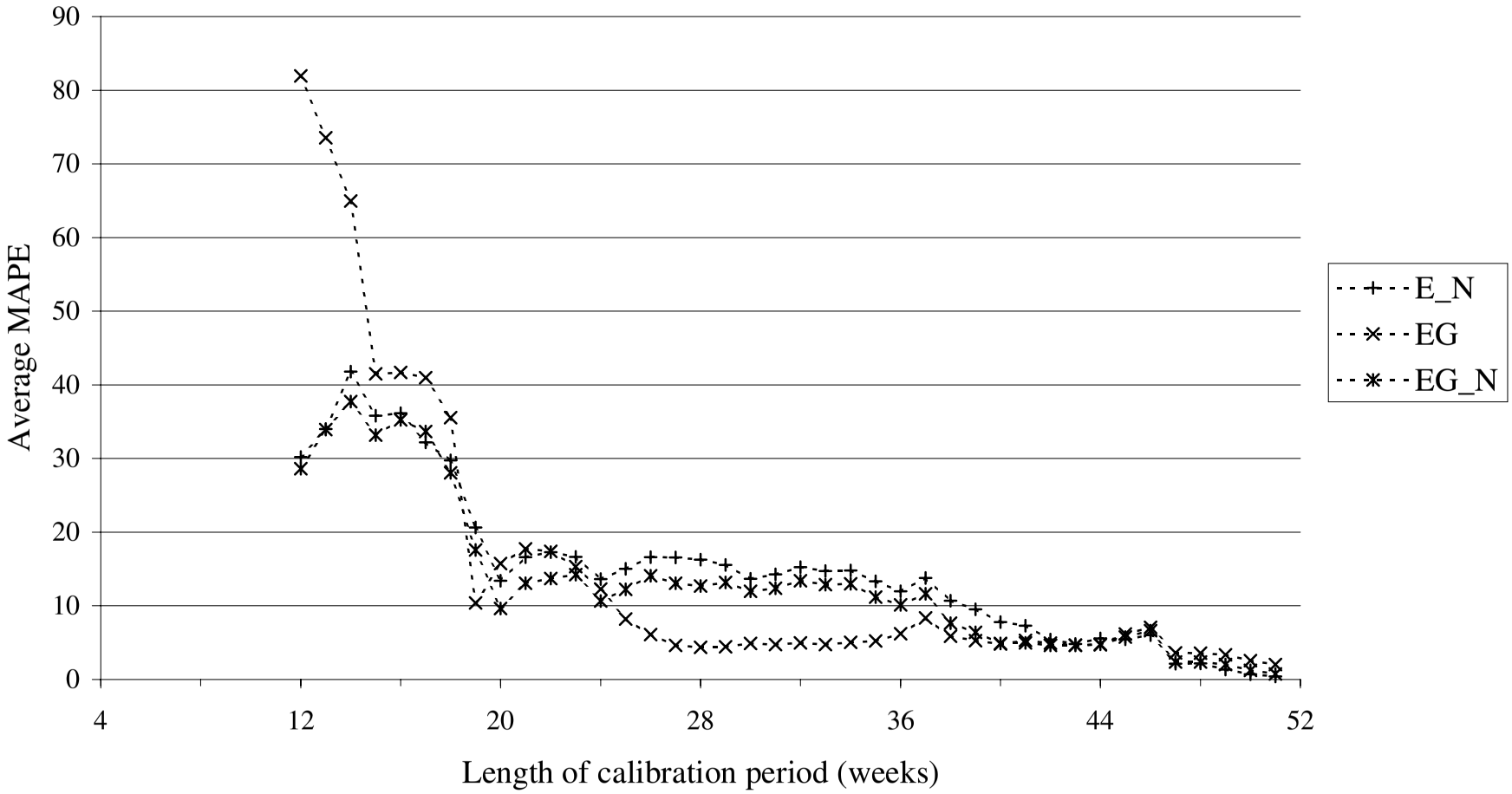

Figure 4:
Forecasting errors: Exponential-gamma models

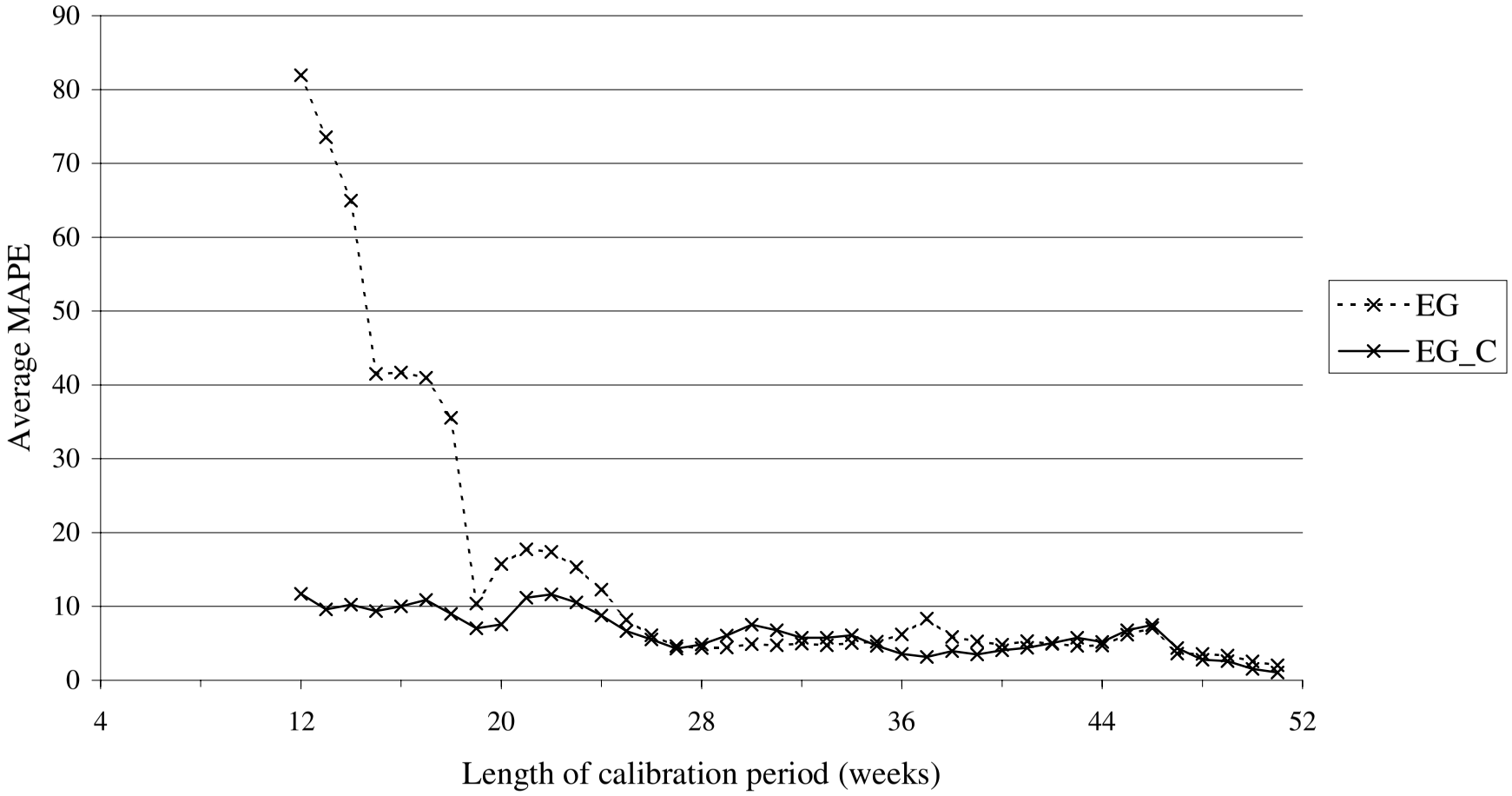

Figure 5:
Parameter stability: Mean time-to-purchase parameter

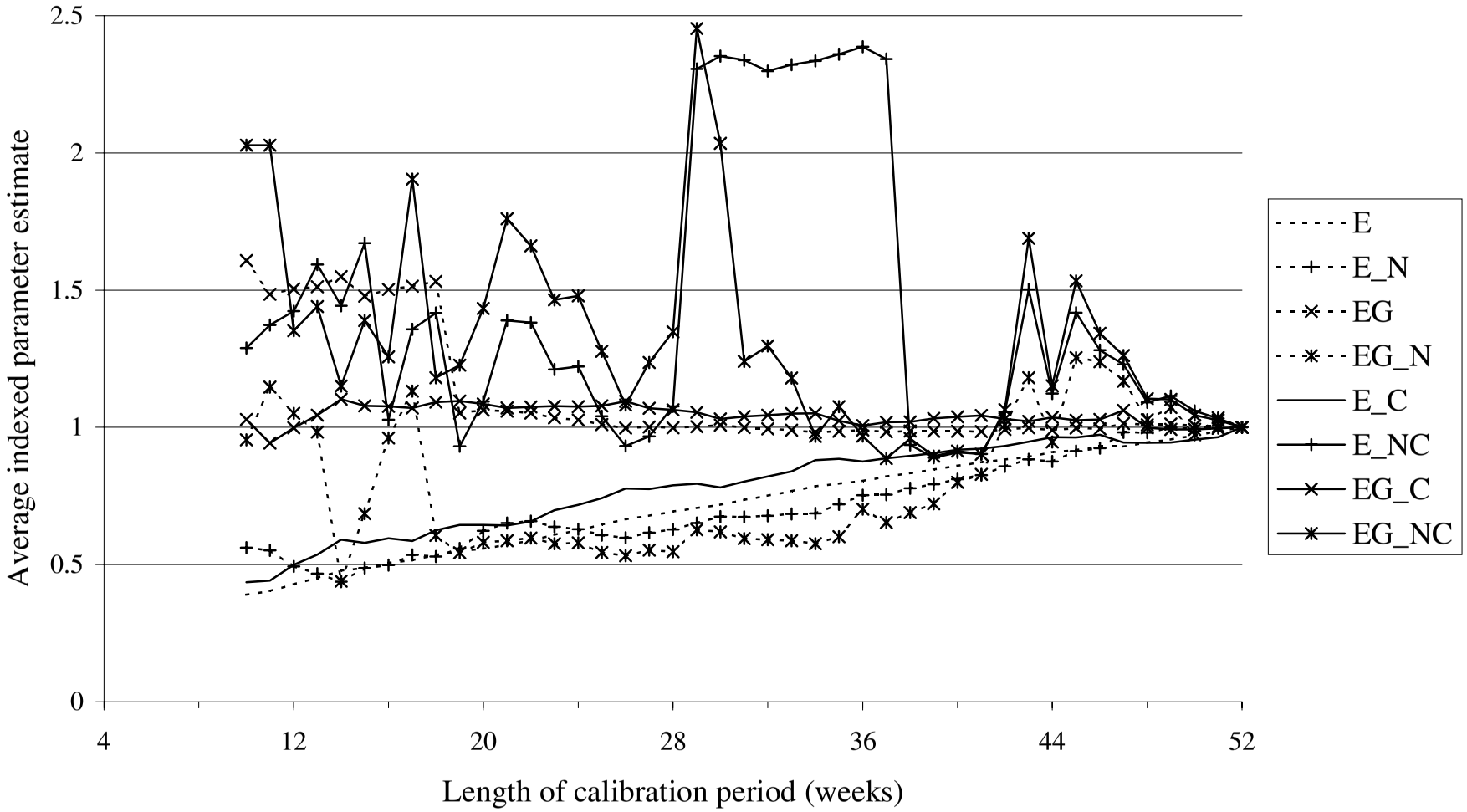

Figure 6:
Parameter stability: Promotion parameter

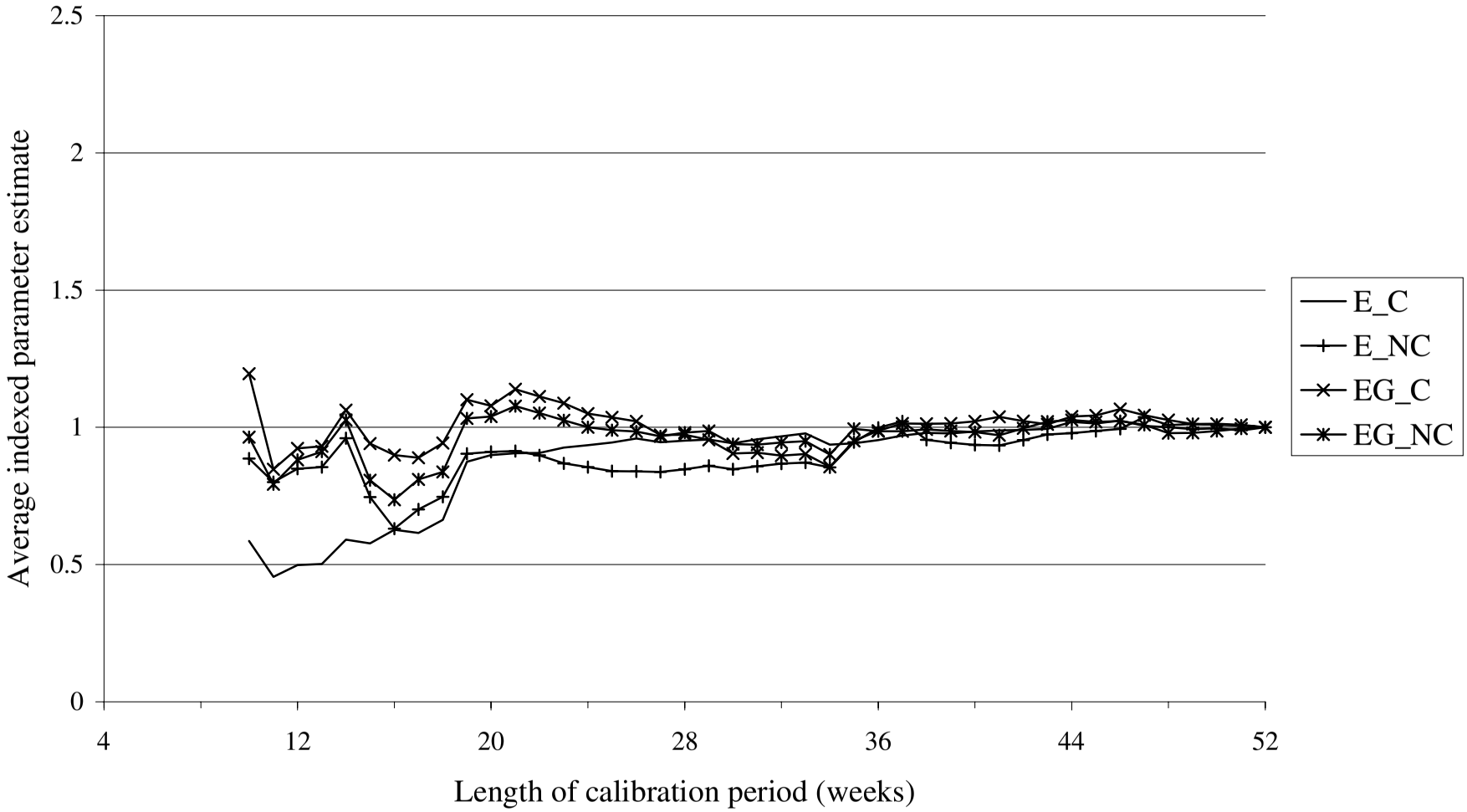

Figure 7:
Parameter stability: Gamma distribution variance parameter