Journal of Forecasting
J. Forecast. **17**, 209–229 (1998)

An Empirical Comparison of New Product Trial Forecasting Models

BRUCE G. S. HARDIE,[1]* PETER S. FADER[2]
and MICHAEL WISNIEWSKI[1]
[1]*London Business School, UK*
[2]*The Wharton School, University of Pennsylvania, USA*

ABSTRACT

While numerous researchers have proposed different models to forecast trial sales for new products, there is little systematic understanding about which of these models works best, and under what circumstances these findings change. In this paper, we provide a comprehensive investigation of eight leading published models and three different parameter estimation methods. Across 19 different datasets encompassing a variety of consumer packaged goods, we observe several systematic patterns that link differences in model specification and estimation to forecasting accuracy. Major findings include the following observations: (1) when dealing with consumer packaged goods, simple models that allow for relatively limited flexibility (e.g. no S-shaped curves) in the calibration period provide significantly better forecasts than more complex specifications; (2) models that explicitly accommodate heterogeneity in purchasing rates across consumers tend to offer better forecasts than those that do not; and (3) maximum likelihood estimation appears to offer more accurate and stable forecasts than nonlinear least squares. We elaborate on these and other findings, and offer suggested directions for future research in this area. © 1998 John Wiley & Sons, Ltd.

KEY WORDS new product forecasting; new product trial; test market

Almost every textbook discussion of the new product development process includes a call to conduct some form of market test before actually launching the new product. Such an exercise serves several objectives, including the desire to produce an accurate forecast of the new product's sales performance over time. These forecasts can help lead to a final go/no-go decision and can also assist in the marketing and production planning activities associated with the product launch. In the case of consumer packaged goods, conducting a market test historically saw the company's sales force selling the product into retail distribution in one or more markets for one to two years, after which a decision of whether or not to go national with the new product was

* Correspondence to: Bruce G. S. Hardie, London Business School, Sussex Place, Regent's Park, London NW1 4SA, UK.

CCC 0277–6693/98/030209–21$17.50

made. Given the problems associated with such traditional ('sell-in') test markets (e.g. cost, time required, competitor sabotage), academics and marketing research practitioners have developed a number of pre-test and test market models and methodologies, all designed to provide the marketing manager with the desired information in a more timely and cost-effective manner.[1] In the case of a test market model, the primary idea is to place the new product in some form of test market and after, say, six months, generate a one- to two-year sales forecast.

The sales volume of a new product in any consumer packaged goods category can be decomposed into *trial* and *repeat* components. The idea behind a test market forecasting model is to develop mathematical representations of these components of sales, and calibrate them using panel and/or survey data. These individual models can then be used to forecast trial and repeat sales, from which an overall forecast of the new product's sales can be generated.

Since 1960, a number of academics and practitioners have proposed various models for the trial and repeat components of a new product's sales. Typically, the developers of a model will provide forecasts for one or two new products, but they rarely compare the performance of their proposed model with that of other existing models. Although there have been several non-empirical comparisons of test market models (e.g. Narasimhan and Sen, 1983), there has been only one empirical comparison of competing models, that of Mahajan, Muller, and Sharma (1984), which examined the awareness components of five models.[2] Motivated by Mahajan and Wind's (1988) call for an 'M-competition' (Makridakis *et al.*, 1982) for new product sales forecasting models, we undertake a study of the relative performance of a set of models of the trial component of new product sales.

Aside from being a logical 'next-step' given the work of Mahajan *et al.*, we focus on trial for four key reasons. First, the accurate forecasting of trial sales is a necessary condition for reliable estimates of the new product's overall sales, as estimates of repeat sales are conditional upon estimates of trial. Even with a perfect estimate of repeat buying rates, over- or underpredicting trial sales will have a corresponding impact on the total sales estimate and can lead to inappropriate go/no-go management decisions. It is therefore clearly important to identify the 'best' trial model (or set of models). Second, product trial is an extremely important diagnostic measure to gauge the short-term status of a new product in the marketplace, as well as an indicator of its long-term success. While low trial rates can often be corrected through appropriate marketing actions, the failure to anticipate them (i.e. waiting for the actual results to occur rather than predicting them using models along the lines of those presented in this paper) could signify lost sales opportunities. Third, identification of the 'best' model(s) results in a benchmark against which future developers of new product forecasting models can compare the performance of their new trial submodels. Finally, by undertaking studies of the relative performance of the various submodels of the trial and repeat components of a new product's sales, there is scope to create 'hybrid' models of total sales which forecast with increased accuracy by combining the best submodels for each component; clearly, this study of trial models is central to such an effort.

The objective of a trial forecasting model is to forecast $P(t)$, the new product's penetration (or cumulative trial) up to some point in time, t. (This is typically measured as the percentage of the panel that has tried the new product by time t, or 'triers per 100 households'.) The data usually used to calibrate these models is simply a time series giving the cumulative numbers

[1] See Clancy, Shulman, and Wolf (1994) and Urban (1993) for recent reviews of pre-test market models and practices, and Narasimhan and Sen (1983) for a review of test market models.

[2] Contrast this to the diffusion/growth modelling literature, in which there have been several studies examining the relative performance of various models (e.g. Meade, 1984; Rao, 1985).

of triers by the end of each time period (e.g. week); no competitive or marketing mix data are used. The simplest and best-known trial model, that originally proposed by Fourt and Woodlock (1960), is based on very simple underlying assumptions of buyer behaviour. While recently developed models may be based on more realistic assumptions, an unanswered question is whether they are correspondingly more accurate in their forecasting performance.

The paper proceeds as follows. First we review the set of trial models whose relative forecasting performance is the focus of our study. We then describe the data used in this study, and the methods used to calibrate the models and determine their relative performance. The forecasting performance of these models is then examined and conclusions are drawn as to which are the more accurate model specifications. We finish with a discussion of a number of issues that arise from our study, and identify several areas worthy of follow-on research.

MODEL REVIEW

The primary criterion used to select the models included in this study is that they have been proposed as models of the trial process for new consumer packaged goods products. Two secondary criteria are used to determine the final set of models examined in this comparative study. The first concerns the source of data used for model calibration. As previously alluded to, test market models can be characterized in terms of whether they utilize panel and/or survey data (Narasimhan and Sen, 1983). The 'decision-process' models (Wilson and Pringle, 1982) that require some survey data—for example, SPRINTER (Urban, 1970), TRACKER (Blattberg and Golanty, 1978), and NEWS (Pringle, Wilson, and Brody, 1982)—are the exception, rather than the rule. As they have different data requirements, they are deemed to be beyond the scope of this study.

The second of the two additional criteria concerns the inclusion of marketing decision variables in the model. Only two researchers—Nakanishi (1973) and Eskin (1974)—have developed panel data-based models that explicitly incorporate such variables, and these are limited in the choice of variables considered (e.g. Eskin considers only advertising). Therefore, in order to facilitate the comparison process, we focus on those models that do not include marketing decision variables. If we wished to comprehensively examine the effects of marketing decision variables, we would have to develop new models by extending the various specifications examined below (cf. the literature on incorporating price and advertising in the Bass model). This, however, is not the purpose of our paper. By excluding decision variables, we are identifying benchmark models against which more complete, but yet to be developed, models can be compared when evaluating the impact of adding covariates. Moreover, while covariate information is available from the panel data source used in this study, not all panel data sources are as rich (in both the USA and the rest of the world). If we were to focus on models with covariate effects, we would be greatly limiting the relevance of the study as the results would be of little use to researchers using many other panel data sources.

On the basis of the above criteria, eight trial models were identified. These models can be characterized in terms of whether they are derived from a set of behavioural assumptions or are simply flexible functional forms designed to 'best fit the data'. The first group of models can be further characterized in terms of: (1) the underlying assumptions made about the probability distribution used to describe a panellist's buying behaviour, (2) whether or not heterogeneity in panellist buying rates is modelled, and (3) whether or not the existence of a group of panellists

classified as 'never triers' is explicitly acknowledged. As most of the selected models can be conveniently described in terms of their underlying behavioural assumptions, our default is to refer to each model in terms of these assumptions, rather than use the developers' names.

Exponential with 'Never Triers'

The best-known model of trial sales is that proposed by Fourt and Woodlock (1960). In examining numerous cumulative trial curves, they noted that (1) successive increments in cumulative trial declined, and (2) the cumulative curve approached a penetration limit of less than 100% of the households in the panel. They proposed that incremental trial 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. This simple model captures the two observed properties.

A continuous-time analogue of this model can be derived from the following two assumptions:

(1) A randomly chosen panellist's time to trial (assuming the new product is launched at time $t = 0$) is distributed according to the exponential distribution; i.e. $f(t) = \theta \exp(-\theta t)$. This is equivalent to saying that a panellist's instantaneous rate of trial, given that he has not yet done so by time t (i.e. $f(t)/[1 - F(t)]$), is a constant, θ.
(2) θ is distributed across the population with the following discrete mixing distribution: $g(\theta) = \lambda$, with probability p; $g(\theta) = 0$, with probability $1 - p$. This captures the fact that some people will simply not be in the market for the new product, i.e. 'never triers'. For example, one would typically expect that diapers will not be purchased by panellists who do not have children (or grandchildren) under 4 years old.

Taken together, these two assumptions result in the following model for the cumulative trial curve:

$$P(t) = p(1 - e^{-\lambda t}) \tag{1}$$

where $P(t)$ is the new product's penetration (i.e. cumulative trial) at time t. This basic model, whether in its original discrete formulation or in the above continuous-time formulation, has served as the starting point for many new product forecasting modeling efforts (e.g. Eskin, 1973; Eskin and Malec, 1976; Parfitt and Collins, 1968).

Exponential with 'Never Triers' + 'Stretch' Factor

Fourt and Woodlock (1960) found that predictions based on the standard exponential with 'never triers' model tend to be too low for later time periods (i.e. the empirical trial curve does not flatten off as quickly as the model predicts). This phenomenon was attributed to heterogeneity in consumer buying rates: heavy category buyers are likely to be the earlier triers and the model picks up the 'levelling out' of their purchases, ignoring the lighter buyers who have yet to try the product (but eventually will).

The proposed solution to this problem was to include a linear 'stretch' factor which allowed the cumulative trial ceiling to be a linear function of time, rather than a fixed quantity. This results in the following model for the cumulative trial curve:

$$P(t) = p(1 - e^{-\lambda t}) + \delta t \tag{2}$$

Fourt and Woodlock (1960), as well as several later researchers (e.g. Eskin, 1973; Kalwani and Silk, 1980), report that δ tends to be a very small positive number, but still plays a significant role in ensuring that the model fits reasonably well.

Exponential-Gamma

An alternative approach to capturing the effect of heterogeneity in consumer buying rates was suggested by Anscombe (1961) and used again by Kalwani and Silk (1980). In particular, Anscombe proposed that this heterogeneity be captured directly in the model. Rather than assuming that the purchasing rate parameter is distributed according to a simple discrete distribution, we can assume that it 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, and $E[\theta] = r/\alpha$. This results in the following model for the cumulative trial curve:

$$P(t) = 1 - \left(\frac{\alpha}{\alpha + t}\right)^r \tag{3}$$

Compared to the use of a 'stretch' factor, the explicit modelling of consumer heterogeneity via a gamma distribution would appear to be a more flexible solution, as well as being more parsimonious than the latter Fourt–Woodlock specification. Given the fact that the inter-arrival times for Poisson events are exponentially distributed, this model can be viewed as the waiting time analogue of the familiar NBD model (Ehrenberg, 1959; Morrison and Schmittlein, 1988).

Exponential-Gamma with 'Never Triers'

While the exponential-gamma model is attractive in that it allows for general types of consumer heterogeneity, it does not explicitly allow for the fact that some people will never try the new product.[3] A natural refinement, therefore, is to utilize both the gamma distribution (to capture heterogeneity among the eventual triers) as well as the p parameter (to screen out the never triers). The resulting model for the cumulative trial curve is as follows:[4]

$$P(t) = p\left\{1 - \left(\frac{\alpha}{\alpha + t}\right)^r\right\} \tag{4}$$

Weibull-Gamma with 'Never Triers'

STEAM, an acronym for STochastic Evolutionary Adoption Model, is a stochastic depth-of-repeat model developed by Massy (see Massy, 1968, 1969; Massy, Montgomery, and Morrison, 1970). Compared to most other new product forecasting models, it has an extremely complex

[3] Implicitly, the gamma mixing distribution can accommodate some of these 'never triers' since it can be highly right-skewed (i.e. for a large proportion of the panelists, θ is very close to zero, implying that their likelihood of trying the new product is very low). A problem with such an approach is that it hinders the interpretation of the mixing distribution's moments (e.g. mean and variance).

[4] This model was examined by Kalwani and Silk (1980) in their study of the structure of repeat buying, and was found to provide a fit superior to the exponential-gamma model. Following Eskin's (1973) practice of applying the same model to trial and all repeat levels, it is natural for us to examine the performance of this model in the context of trial.

structure. However, it is an *integrated* model, thereby alleviating the need to combine the results of separate models (e.g. trial, first repeat, additional repeat) when generating a sales forecast for the new product.

At the heart of the STEAM model is a relaxation of the exponential assumption that a panellist's instantaneous rate of trial, given that he has not yet done so by time t, is a constant. Rather, it is a function of time (i.e. interpurchase times are assumed to follow a form of the Weibull distribution). Heterogeneity in base purchase rates is captured via a gamma mixing distribution, giving us a form of compound Weibull distribution. The STEAM model for trial sales is developed from this distribution by making several additional assumptions (e.g. the gamma mixing distribution is only applied to those panellists who will eventually try the new product), resulting in the following model for the cumulative trial curve:

$$P(t) = p\left\{1 - \left[\frac{\alpha c}{(t+1)^c + \alpha c - 1}\right]^r\right\} \tag{5}$$

where c is the shape parameter of the Weibull distribution. When $c = 1$, this reduces to the exponential-gamma with the 'never triers' model.

Lognormal–Lognormal

The models reviewed so far have assumed that a panellist's purchase rate follows an exponential distribution (or one of its generalizations). An alternative view is taken by Lawrence (1979, 1982, 1985), who postulates that interpurchase times at the individual panellist level are lognormally distributed. By assuming that the mean interpurchase times across the panellists are also distributed lognormal, the following model of the cumulative trial curve can be derived by a straightforward application of renewal theory:[5]

$$P(t) = \frac{t}{e^{\mu+\sigma^2/2}}[1 - \Lambda(t \mid \mu, \sigma^2)] + \Lambda(t \mid \mu + \sigma^2, \sigma^2) \tag{6}$$

where $\Lambda(t \mid \mu, \sigma^2)$ is the lognormal distribution function with mean μ and variance σ^2 (i.e. $\ln(t)$ is normally distributed with mean μ and variance σ^2).

'Double-Exponential'

According to Greene (1974, p. 419), 'every new-brand test-marketer knows that cumulative trial usually follows an S-curve'. He then proposes the following model for the cumulative trial curve:

$$P(t) = \frac{p}{\beta - \alpha}[\beta(1 - e^{-\alpha t}) - \alpha(1 - e^{-\beta t})] \tag{7}$$

Unlike the previous models, which can all be derived from simple assumptions about the buying behaviour of panellists, this formulation is simply the result of a curve-fitting exercise.[6] It is worth

[5] The derivations reported in Lawrence (1982, 1985) express the cumulative trial curve in terms of three lognormal distribution parameters—the variance of the interpurchase time lognormal distribution, and the mean and variance of the lognormal mixing distribution. The form of the model presented here is that outlined in Lawrence (1979), where μ equals the mean of the lognormal mixing distribution less the variance of the mixing distribution and half the variance of the panellist-level interpurchase time lognormal distribution, and σ^2 is the sum of the variances of the two lognormal distributions.

[6] While Greene uses the term 'double-exponential' to describe this model, we should not confuse it with the family of double exponential (Laplace and extreme value) probability distributions (see Johnson, Kotz, and Balakrishnan, 1994).

noting that Greene's insistence that cumulative trial curves are S-shaped is contrary to the view implicitly held by many other developers of new product forecasting models.

The Bass Model

Within the marketing literature, the best-known diffusion model is that developed by Bass (1969). The central premise of this model is that the probability of adopting (or, in this context, trying) the product at time t, given that adoption (trial) has not yet occurred, equals $\alpha + \beta \times$ cumulative proportion of adopters (triers) at time t. This, combined with the added assumption that only a proportion p of the population will ever adopt (try) the new product, results in the following model for the cumulative adoption (trial) curve:

$$P(t) = p\left[\frac{1 - e^{-(\alpha+\beta)t}}{1 + (\beta/\alpha)\, e^{-(\alpha+\beta)t}}\right] \tag{8}$$

(Note that this collapses to the exponential with the 'never triers' model when $\beta = 0$.)

While this model has been used countless times to forecast the sales of durables, it has been used *very* rarely in the context of consumer packaged goods. As the literature includes at least one example (Burger, 1968), the model meets the primary criterion for consideration in this study.[7]

Other Models

Table I summarizes the characteristics of the above set of models and demonstrates the broad coverage of the main classes of model in the literature. As previously noted, the primary criterion used to select the models included in this study is that they have been proposed as models of the trial process for new consumer packaged goods products. With one exception, we have therefore disregarded the various 'diffusion' models widely examined in the marketing literature. (See Mahajan, Muller, and Bass, 1990, and Parker, 1994 for reviews of the diffusion modelling literature.) As noted by Parker (1994, p. 356), a basic component of diffusion models is 'a growth rate component which can characterize interpersonal influence among members of the target market'. The theoretical rationale for using such models in our context (trial sales of consumer packaged goods) is weak, as most frequently purchased packaged goods are low-involvement products for which there is usually little risk and low uncertainty associated with trying the new product (Gatignon and Robertson, 1985) and for which we would expect to see minimal 'word-of-mouth' effects. (When did the reader last tell a friend about that new brand of toilet tissue he just purchased?) The opposite is typically the case for the types of products to which diffusion models are applied (e.g. durables).

Given that we do not anticipate seeing 'word-of-mouth' effects for consumer packaged goods products, what could explain an observed S-shaped cumulative trial curve? Jones and Mason (1990) demonstrate that it might be changes (i.e. growth) in retail distribution in the first few weeks after product launch. As test market models are frequently applied in a controlled distribution environment, there is even less reason to expect to see S-shaped cumulative trial curves. In those cases where distribution is not controlled, the S-shape should be captured via the appropriate specification of a distribution effect, *not* via a flexible functional form where the extra

[7] The model has also been utilized in a consumer packaged goods setting for descriptive, rather than forecasting, purposes by Parker and Gatignon (1994).

Table I. Summary of model characteristics

Model	Equation	Structural model	Heterogeneity	'Never triers'
1	$P(t) = p(1 - \mathrm{e}^{-\lambda t})$	Exponential	None	✓
2	$P(t) = p(1 - \mathrm{e}^{-\lambda t}) + \delta t$	Exponential	Heuristic correction term	✓
3	$P(t) = 1 - \left(\frac{\alpha}{\alpha + t}\right)^r$	Exponential	Gamma	
4	$P(t) = p\left\{1 - \left(\frac{\alpha}{\alpha + t}\right)^r\right\}$	Exponential	Gamma	✓
5	$P(t) = p\left\{1 - \left[\frac{\alpha c}{(t+1)^c + \alpha c - 1}\right]^r\right\}$	Weibull	Gamma	✓
6	$P(t) = \frac{t}{\mathrm{e}^{\mu + \sigma^2/2}}[1 - \Lambda(t \mid \mu, \sigma^2)] + \Lambda(t \mid \mu + \sigma^2, \sigma^2)$	Lognormal	Lognormal	
7	$P(t) = \frac{p}{\beta - \alpha}[\beta(1 - \mathrm{e}^{-\alpha t}) - \alpha(1 - \mathrm{e}^{-\beta t})]$	None	None	✓
8	$P(t) = p\left[\frac{1 - \mathrm{e}^{-(\alpha+\beta)t}}{1 + (\beta/\alpha)\,\mathrm{e}^{-(\alpha+\beta)t}}\right]$	Custom	None	✓

parameter is given a behavioural interpretation unrelated to the true causes that affect its estimated value.

We have included four models that allow for an S-shaped cumulative trial curve. These include the Bass and 'Double-Exponential' models, as already discussed, but can also include the Weibull-gamma with 'never triers' and lognormal–lognormal models. These last two models do not always generate an S-shape but can do so via the combination of a non-constant hazard rate at the household level and a skewed heterogeneity distribution. Other possible S-curve models would be the various logistic models reviewed in Massy, Montgomery, and Morrison (1970, Chapter 8); these models are not considered as they have not been explicitly used as trial forecasting models in a consumer packaged goods setting. Another class of models not examined in this study, for the same reason, are those based on linear learning assumptions (Massy, Montgomery, and Morrison, 1970, Chapter 8). Finally, as we are focusing on models developed for new product forecasting, we have not included any models of buying behaviour for established products that can also be applied to new products (e.g. Aaker, 1971; Herniter, 1971).

EMPIRICAL ANALYSIS

The data used to estimate trial models of the type included in this study are derived from consumer panels. At the time these models were developed, panel data was typically collected via self-completed diaries. With the adoption of the Universal Product Code (UPC) and associated laser scanners, diaries have been replaced by some form of electronic data collection, be it in the store (where panellists present a special ID card to the cashier at the checkout) or in the home (where panellists scan the items purchased using a hand-held barcode scanner).

The specific data used in this study are from market tests conducted using Information Resources, Inc.'s *BehaviorScan* service. *BehaviorScan* is an electronic test marketing system with panels operating in eight markets, geographically dispersed across the USA; 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). One characteristic of this service is that distribution is controlled; i.e. the new product receives 100% distribution in the market. (See Curry, 1993 for further details of the *BehaviorScan* service.)

We have 19 datasets, each associated with a new product test (lasting one year) conducted in one of the targetable TV markets between 1989 and 1996. Each of the six markets is represented at least once in our database, and the tested products are from the following categories: shelf-stable (ready-to-drink) juices, cookies, salty snacks, and salad dressings. (Further details cannot be provided for reasons of confidentiality.) It should be noted that some new products were simultaneously tested in two markets; in such cases, we treat each market as a separate observation (i.e. fewer than 19 unique new products were tested).

Each product-market dataset contains 52 weekly observations, each observation being the cumulative number of panellists who have tried the new product by the end of the week in question. These are converted to penetration numbers simply by dividing by the number of households in the panel.

Traditional industry practice sees forecasts being made once the product has been in the test market for six months. We therefore use the first 26 weeks of data to calibrate each trial model, and the remaining 26 weeks of data are used for evaluating the model's forecasting performance. It is clear that getting an earlier read of the new product's performance is of great interest to both the research companies and their clients, and the desire to use a 13-week test period is often discussed. We therefore consider a second calibration period condition in which the first 13 weeks of data are used for model calibration and the remaining 39 weeks of data for evaluating forecasting performance.

Model Calibration

A number of the original developers of the models included in this study provided little or no detail as to how they calibrated the models for their empirical analyses. Following the lead of Mahajan, Mason, and Srinivasan (1986), we will examine the impact of different estimation methods—maximum likelihood (MLE) and two variants of non-linear least squares (NLS)—on the predictive validity of the eight models included in this review. The only exception is the exponential with the 'never triers' + 'stretch' factor model, which, due to the inclusion of the stretch factor (δ) parameter, is not amenable to MLE. In total, therefore, we have a set of 23 different model × estimation method combinations.

Let:

$Y(t)$ = the new product's actual penetration by the end of week t
$y(t)$ $= Y(t) - Y(t-1)$
= the new product's incremental trial in week t
$p(t)$ $= P(t) - P(t-1)$
= the model-based estimate of incremental trial in week t
N = the number of households in the panel
T = the number of weeks in the calibration period.

Maximum likelihood estimates of each model's parameters are obtained by maximizing the following general likelihood equation; the exact equation is derived by substituting in the specific expressions for $P(t)$ and $p(t)$:

$$L = [1 - P(T)]^{N[1-Y(T)]} \prod_{t=1}^{T} p(t)^{Ny(t)}$$

In the diffusion and growth modelling literature, researchers tend to take one of two approaches in using non-linear least squares to estimate a model's parameters. Srinivasan and Mason (1986) obtain parameter estimates by minimizing the sum-of-squares on an incremental (i.e. week-by-week) basis:

$$\sum_{t=1}^{T} [y(t) - p(t)]^2$$

while Meade (1984) obtains parameter estimates by minimizing the sum-of-squares based on the number of cumulative triers through time t:

$$\sum_{t=1}^{T} [Y(t) - P(t)]^2$$

We will use both approaches, calling them NLS-Incremental and NLS-Cumulative, respectively.

Measurement of Forecasting Performance

Using the parameters estimated on the first 13 (or 26) weeks of data, each model is used to forecast cumulative trial for each of the remaining 39 (or 26) weeks. In summarizing a model's ability to forecast cumulative trial, we are first interested in week-by-week accuracy, as repeat sales forecasts derived from models of the repeat component of sales are typically computed on a week-by-week basis, conditional on cumulative trial. Failing to appropriately track cumulative trial would ripple though to the eventual total (i.e. trial + repeat) sales forecast. Another summary measure frequently used in market tests is cumulative trial by the end of the first year. We are therefore also interested in the models' ability to forecast cumulative trial at week 52.

The appropriate error measures will be computed for each of the 874 model × estimation method × calibration period × dataset combinations. These will then be analysed to identify the relative performance of the eight models, and the impact of the three estimation methods and two calibration period lengths. 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. A commonly used measure is Mean Absolute Percentage Error (MAPE). While there are theoretical advantages associated with the use of alternative measures—see Armstrong and Collopy (1992), Fildes (1992), and related commentary—MAPE has the advantages of being easily understood by managers and being very appropriate in planning and budgeting situations (Makridakis, 1993). We will therefore compute MAPE over the forecast period for each of the model × estimation method × calibration period × dataset combinations. Week 52 forecasting accuracy will be evaluated by computing the Absolute Percentage Error for that week alone (hereafter APE_52).

Table II. Average calibration period *R*-squared

Model	Calibration period = 13 weeks				Calibration period = 26 weeks				Overall mean[b]
	MLE	NLS-Cum	NLS-Inc	Mean[a]	MLE	NLS-Cum	NLS-Inc	Mean	
1. Exp. w/NT	0.968	0.967	0.958	0.964	0.960	0.965	0.958	0.961	0.963
2. Exp. w/NT + 'stretch'	—	0.968	0.959	0.963	—	0.971	0.960	0.965	0.964
3. Exp.-Gamma	0.950	0.956	0.944	0.950	0.946	0.954	0.937	0.946	0.948
4. Exp.-gamma w/NT	0.968	0.967	0.958	0.964	0.963	0.967	0.958	0.963	0.964
5. Weib.-gamma w/NT	0.982	0.983	0.973	0.979	0.971	0.974	0.964	0.970	0.975
6. Lognormal–lognormal	0.937	0.945	0.931	0.938	0.924	0.938	0.914	0.925	0.932
7. 'Double-exponential'	0.974	0.979	0.977	0.977	0.961	0.967	0.956	0.961	0.969
8. Bass	0.970	0.979	0.968	0.972	0.960	0.966	0.925	0.950	0.961
Mean[c]	0.964	0.968	0.959		0.955	0.963	0.946		

[a] Model average across estimation method.
[b] Model average across estimation method and calibration period.
[c] Estimation method average across model.

RESULTS

To help set the stage for the forecasting results that follow, we first examine the relative performance of each model and estimation method within the two calibration periods (13 and 26 weeks). Table II contains the summary results. (Note that the numbers in the left-hand column of Table II have nothing to do with the relative performances of the models; they merely reflect the ordering of the models as presented in the second section.) The models all fit quite well in calibration, and, in general, there are no critical differences across the model × estimation method × calibration period combinations. We briefly summarize the few noteworthy differences that do emerge from the table:

(1) As judged by the R^2 criterion, the two models lacking a p parameter (to allow for ultimate penetrations less than 100%)—namely, exponential-gamma (model 3) and lognormal–lognormal (model 6)—both fit noticeably worse than the other six specifications. This observation holds true not only for the overall means (final column of Table II), but for each of the different estimation method × calibration period combinations as well.
(2) Regardless of the calibration interval, NLS-Cum offers slightly better fits than MLE, which, in turn, fits slightly better than NLS-Inc. The better fits provided by NLS-Cum come as no surprise, since such an optimization process is essentially equivalent to maximizing the R^2 for a given model.
(3) The model × estimation method combinations using 13 weeks of calibration data tend to produce slightly higher R^2s than those using 26 weeks. Most likely, these differences reflect some degree of overfitting for the models run with shorter calibration periods.

Beyond these main effects, there are no apparent interactions occurring for the various combinations of models, estimation methods, and calibration periods. The more important issue, of course, is whether the presence of any of these calibration period effects are associated with any meaningful (or at least statistically significant) differences in the forecasts. We will now address this essential question.

Table III. Average forecast period MAPE

Model	Calibration period = 13 weeks				Calibration period = 26 weeks				Overall mean[b]
	MLE	NLS-Cum	NLS-Inc	Mean[a]	MLE	NLS-Cum	NLS-Inc	Mean	
1. Exp. w/NT	18.31	15.58	15.31	16.40	8.81	12.15	9.57	10.17	13.29
2. Exp. w/NT + 'stretch'	—	22.22	20.75	21.49	—	6.76	6.17	6.47	13.98
3. Exp.-gamma	17.91	17.08	20.71	18.56	5.89	9.80	12.42	9.37	13.97
4. Exp.-gamma w/NT	21.38	15.63	16.26	17.76	7.29	8.89	7.12	7.77	12.76
5. Weib.-gamma w/NT	19.32	19.87	21.51	20.23	7.41	8.96	11.65	9.34	14.78
6. Lognormal–lognormal	26.96	26.89	34.38	29.41	9.75	13.55	21.41	14.90	22.16
7. 'Double-exponential'	21.42	21.21	20.48	21.04	9.02	13.03	14.55	12.20	16.62
8. Bass	18.50	22.40	25.69	22.20	8.26	13.37	22.51	14.71	18.45
Mean[c]	20.54	20.11	21.88		8.06	10.81	13.17		

[a–c] As Table II.

MAPE Results

The summary data for our MAPE analyses can be found in Table III. A number of important patterns emerge from this performance measure. We first discuss the MAPE results associated with each of the eight model specifications, and then broaden our focus to examine issues involving the calibration periods and estimation methods as well.

Our principal result can be seen in the rightmost column of Table III. The MAPE numbers here reflect a strong pattern favouring simpler models over more complex ones. Specifically, the top four model specifications all offer better forecasts on average (i.e. lower MAPEs) than the lower four models. The qualitative differences across these two foursomes is quite striking: the top four models are purely concave; under no circumstances can any of them allow for any 'S-shaped' behaviour. In contrast, the lower four models all allow for some degree of convexity. It is evident from these results that more flexibility in a model specification is not necessarily a good thing, and, in fact, it can be downright harmful to a model's forecasting capabilities. Overall, these differences in mean MAPE are highly significant ($F_{7,866} = 7.385$, $p < 0.001$), yet the differences among the top four models are not at all significant ($F_{3,866} = 0.234$, $p = 0.834$). Thus the simple, concave models are, in aggregate, very similar to each other (in terms of forecasting performance), but substantially more accurate than the four complex structural models. This is perhaps the most important finding of this entire study.

Beyond these overall main effects for the eight models, a number of deeper insights emerge when we examine the model-by-model results under the various estimation methods and calibration periods. For instance, when we look closely at the model results associated with the 13-week calibration period, we see the same overall pattern emerging (i.e. simpler models generally forecasting better than complex ones), but there is a notable exception. The second model, exponential with the 'never triers' + 'stretch' factor, falls far behind the other three simple models, and in fact is surpassed by two of the complex models. (Nevertheless, the differences among the other three simple models continue to be insignificant: $F_{2,429} = 0.371$, $p = 0.690$.)

In sharp contrast, this same model (exponential with the 'never triers' + 'stretch' factor) is the clear winner for the 26-week calibration period. To add to this surprising result, the exponential with the 'never triers' model *without* the 'stretch' factor (model 1), which was the overall winner for 13 weeks, drops to fifth place when the longer calibration period is used. Clearly, this 'stretch'

factor becomes critically important in longer datasets. Recall that Fourt and Woodlock (1960) introduced this parameter to accommodate unobserved heterogeneity to enhance the performance of the simple exponential model. Our results suggest that this stretch factor does its job extremely well, but only when the data period is sufficiently long so as to make heterogeneity an important consideration.

Despite these flip-flops for the exponential model with and without the 'stretch' factor, the two exponential-gamma specifications (models 3 and 4) are quite steady and accurate across the two calibration periods. This robustness is a very encouraging sign, suggesting that these models may be broadly applicable across a relatively wide variety of datasets.

Among the complex models, the lognormal–lognormal (model 6) is consistently the worst performer, and by a wide margin. (Recall that it was also the single worst performer in terms of calibration period R^2.) The Bass model (model 8) is next-worst, with the 'Double-Exponential' (model 7) just ahead of it. The Weibull-gamma with 'never triers' (model 5) is clearly the best of the four complex models. Ironically, it finishes in fourth place within each of the two calibration periods, yet in fifth place overall (due to the wild swings of models 1 and 2).

MAPE by Estimation Method

The bottom row of Table III shows the mean MAPE for each of the three estimation methods, broken out by calibration period. If we aggregate these two groupings together, we get the following trio of mean MAPE values: 14.302 for MLE, 15.461 for NLS-Cum, and 17.529 for NLS-Inc. The difference across these means is highly significant ($F_{2,871} = 4.766$, $p = 0.009$), but this is driven largely by the fact that NLS-Inc is clearly the worst estimation method (overall as well as in each of the two calibration periods). In contrast, MLE and NLS-Cum are somewhat harder to distinguish from one another. Overall, the difference between MLE and NLS-Cum is insignificant ($F_{1,871} = 1.180$, $p = 0.278$) when aggregated across the calibration periods. While this insignificant difference between MLE and NLS-Cum holds within the 13-week period ($F_{1,434} = 0.068$, $p = 0.794$), the difference for 26 weeks is quite significant ($F_{1,434} = 7.051$, $p = 0.008$). In other words, NLS-Cum appears to hold a slight (but insignificant) edge for short calibration periods, but the forecasts associated with MLE become stronger and stronger as the calibration period lengthens.

Finally, there appear to be a few interesting model × estimation method interactions worth noting. First of all, the simple exponential-gamma (model 3) seems to work very well with MLE. For both calibration periods, this is the best model associated with MLE. Similarly, the Bass model does unusually well with this estimation method—despite its poor performance overall—when compared to the MLE forecasts for the three other complex models (cf. Mahajan, Mason, and Srinivasan, 1986). In contrast, the other estimation methods do not appear to offer similarly consistent results for the various models. The best model × estimation method combinations differ by calibration period, and therefore do not suggest a high degree of reliability or external validity.

We summarize all these MAPE results as follows. Simple models clearly outforecast the more complex specifications, despite the fact that the latter models can allow for 'S-shaped' sales patterns. Overall, the four simple models appear to offer similar forecasting capabilities, but there are some notable differences based on the length of the calibration period and the estimation method used. We observed that the two exponential-gamma models are generally the most stable, and the overall best combination appears to be the simpler of the two exponential-gamma models (i.e. model 3, with no 'never triers' component), which produces especially good forecasts when

Table IV. Average week 52 APE

Model	Calibration period = 13 weeks				Calibration period = 26 weeks				Overall mean[b]
	MLE	NLS-Cum	NLS-Inc	Mean[a]	MLE	NLS-Cum	NLS-Inc	Mean	
1. Exp. w/NT	28.53	24.26	24.04	25.61	15.00	17.67	15.12	15.93	20.77
2. Exp. w/NT + 'stretch'	—	38.19	35.23	36.71	—	8.76	10.06	9.41	23.06
3. Exp.-gamma	25.88	22.60	28.56	25.68	8.59	11.60	15.44	11.88	18.78
4. Exp.-gamma w/NT	34.38	24.23	24.42	27.68	12.58	11.68	10.61	11.63	19.65
5. Weib.-gamma w/NT	28.59	28.55	30.08	29.07	12.04	11.70	16.13	13.29	21.18
6. Lognormal–lognormal	42.19	39.24	51.98	44.47	15.60	17.26	29.54	20.80	32.63
7. 'Double-exponential'	32.31	32.59	31.93	32.25	23.40	19.03	21.59	21.34	26.79
8. Bass	28.70	33.35	35.75	32.60	14.43	19.53	34.11	22.69	27.64
Mean[c]	31.50	30.38	32.75		14.52	14.65	19.08		

[a–c] As Table II.

estimated using MLE. This particular combination produced forecasts with average MAPEs under 18% using 13 weeks of calibration data, and just below 6% on average for the 26-week calibration period.

APE_52 Results

The second forecasting performance measure of interest is APE_52, which captures the accuracy of the models' forecasts in week 52. The summary data for these analyses are shown in Table IV. For the most part, these numbers closely echo the MAPE results just discussed. Rather than review these patterns in detail, we can summarize the highlights as follows:

(1) The simpler specifications (models 1–4) generally provide more accurate week 52 estimates than the more complex models. The two exponential-gamma models are the best overall.
(2) When we contrast the results across the two calibration periods, we continue to see the same major swings for models 1 and 2 (the exponential with the 'never triers' models).
(3) NLS-Cum forecasts slightly better for shorter calibration periods, but MLE is the winner for longer datasets. NLS-Inc continues to lag far behind.
(4) The exponential-gamma model estimated using MLE over 26 weeks is still the single best model × estimation method combination.

In essence, none of the MAPE results are contradicted or markedly amplified by the APE_52 statistics. This consistency across Tables III and IV is a very encouraging sign: it suggests that the choice of week 52 as a focal point is not a critically important one. Presumably, we could perform a similar APE analysis for a different forecasting horizon (within reason) and expect roughly equivalent results to emerge. To pursue this point a little further, we can examine the differences between the rightmost columns of Tables III and IV as an indication of the degree of degradation that occurs for each model as we shift our focus from the mean of each forecast period (MAPE) towards the endpoint (APE_52). In general, these differences are modest, yet they vary systematically across the eight models. The models with the worst forecasting performances (models 6, 7, and 8) suffer from the most degradation, while exponential-gamma (model 3) has by far the smallest difference between its overall means for MAPE and APE_52. This seems to indicate that model 3 might hold up best if we were to further extend the forecasting horizon. While a formal

version of such an analysis might be an interesting endeavour, it is outside the scope of this paper (and these datasets). Nevertheless, we encourage future researchers to pursue this direction of analysis more carefully.

Impact of the Penetration Limit (p) on Forecasts

Missing from the preceding analysis of APE_52 is any insight as to whether the models exhibit any tendency to over- or underpredict. A recent examination of the properties of macro-level diffusion models by Van den Bulte and Lilien (1997) highlights the tendency of such models to underestimate market size. In the context of the models examined in this paper, this would suggest that the models that have a penetration limit term (i.e. explicitly consider 'never triers') will exhibit a bias to underpredict.

In order to explore this idea, we look at the six models that include a penetration limit term to see whether there is any tendency to over- or underpredict, and relate this to whether or not the actual penetration by week 52 is greater or less than the estimated penetration limit; the associated cross-tab is presented in Table V. We observe a clear tendency to underpredict week 52 penetration, with this occurring in 80% of the cases examined. This is obviously linked to the under-estimation of p; in 86% of these underprediction cases, the estimated penetration limit was less than the actual penetration at week 52.[8] Thus, our data seem to reflect the same bias identified by Van den Bulte and Lilien, although in a different context (i.e. models and datasets) from that used by them. As suggested by Van den Bulte and Lilien, researchers must exercise caution in estimating and interpreting these penetration limit parameters.

A logical follow-on question is whether the models containing no penetration limit term (i.e. models 3 and 6) exhibit any tendency to over- or underpredict. Across all estimation method × calibration period × dataset combinations, the exponential-gamma and lognormal–lognormal models overpredict 74% and 91% of the time, respectively.

This analysis shows that the models do exhibit a tendency to over- or underpredict, and this can be linked to the explicit consideration of 'never triers' in the model formulation. In order to overcome the problem of underestimating p, Van den Bulte and Lilien suggest that researchers

Table V. The impact of estimated penetration limit on week 52 forecast

		Underprediction at week 52 $[P(52) < Y(52)]$?		
		No	Yes	
Estimated penetration limit less than actual week 52 penetration $[\hat{p} < Y(52)]$?	No	17%	11%	28%
	Yes	3%	69%	72%
		20%	80%	

[8] On the surface, it would not seem possible to overpredict at week 52 when the estimated penetration limit is less than actual penetration at week 52. The 3% of cases where this does occur are all for the exponential with 'never triers' + 'stretch' factor model (i.e. model 2); a positive δ term enables the forecast to be greater than the estimated penetration limit.

should use externally obtained estimates of the penetration limit. While this may be feasible for some classes of durables, it is not immediately clear how this would have done in the context examined in this paper as it would require an exogenous estimate of the size of the 'never triers' group. Clearly this is an issue that warrants further attention.

DISCUSSION AND CONCLUSIONS

One of the primary purposes of this study has been to address the question 'which trial model(s) provide the best forecasts?' While we are prepared to answer this question, the more important contributions from this paper arise from the deeper question, 'when and why do these models perform best?' We choose to focus on these fundamental issues first. Several general principles have emerged from our analysis that supersede the relative performance of any one particular model specification. We summarize these principles as follows:

(1) When dealing with consumer packed goods, simpler models offer better forecasts than more complex ones. In particular, the family of models associated with an underlying exponential timing process are considerably better than all alternatives, including those models that allow for S-shaped cumulative sales curves.
(2) Models that explicitly accommodate heterogeneity in purchasing rates across consumers tend to offer better forecasts than those that do not do so.
(3) The presence or absence of an explicit 'never triers' parameter (i.e. a penetration limit) does not appear to have any systematic impact on the overall degree of forecast accuracy. At the same time, however, models that include such a component exhibit a strong tendency to under-forecast actual trial sales, while models without this parameter are very likely to overforecast actual trial sales.
(4) In choosing among the two most common estimation procedures, maximum likelihood is significantly better than week-by-week (incremental) NLS. A less common third alternative, cumulative NLS, appears to be similar to MLE overall, but tails off with longer calibration periods.
(5) Not surprisingly, the length of the calibration period has a major impact on the quality of a model's forecast. More subtle is the fact that certain models (e.g. the exponential with 'never triers' specifications) are highly sensitive to the amount of calibration data available for model estimation.
(6) A model's fit within the calibration period is largely unrelated to its forecasting prowess. (One notable exception is the lognormal–lognormal model, which is extremely poor in both domains.) Likewise, differences across the three estimation methods within the calibration period show no relationship with forecasting capabilities.

Given the number and variety of datasets used here, we have reason to believe that these principles are generalizable to many other consumer packaged goods settings. But we acknowledge that our ability to make such generalizations does not necessarily carry over to industrial products, consumer durables, and other types of product categories. As noted earlier, we expect that the ability to accommodate S-shaped sales patterns can be very useful in some of these other contexts. Nevertheless, the universe of consumer packaged goods is a large and varied one, and, as mentioned in the introduction, there is an important role for accurate, reliable trial forecasts.

Looking across this collective set of principles, we can see why the exponential-gamma models (with or without the 'never triers' component) deserve to be crowned as the overall champions of this forecasting challenge. These models effectively combine simplicity with flexibility (for consumer heterogeneity) and appear to be highly robust across calibration periods and estimation methods. It is unfortunate that these models have received relatively little attention from practitioners and academics alike. We hope that this paper might motivate greater use of this class of models, especially in conjunction with some of the extensions discussed below.

The Role of Covariates

While academics and practitioners are certainly interested in forecast accuracy and the other issues discussed in this paper, they are also vitally concerned about the role that covariates, such as marketing decision variables, might have on a model's forecasting capability. After all, an important aspect of a new product's introduction is its marketing plan, which can greatly affect its first-year sales. (This is especially true for firms that use real or simulated test markets to help fine-tune the eventual full-scale marketing plan.) While we share this interest in exploring the impact of marketing decision variables, we must reiterate our earlier observation that remarkably few published trial models have incorporated such effects, and these have received little attention (e.g. Nakanishi, 1973). Thus, the inclusion of covariates requires the development of new models, which is clearly outside the scope of this paper. By excluding covariates, we have identified benchmark models against which more complete models can be compared when evaluating the impact of adding covariates. Moreover, the 'best' models identified in this paper could be ideal base models to which covariate effects can be added. A useful analogy here may be the so-called 'Generalized Bass Model' (Bass, Krishnan and Jain, 1994), which uses a well-founded approach for including covariate effects and demonstrates the robustness of the basic model specification to the omission of covariates. In light of the evidence from Bass *et al.*, as well as other similar studies (e.g. Morrison and Schmittlein, 1988), we have no reason to believe that any of our general conclusions about the models, estimation methods, etc. would change if covariates were included. Nevertheless, we leave this as a worthwhile empirical issue to be addressed more carefully in the future.

Repeat Purchase Models

Despite the importance of modelling trial, *per se*, in understanding and forecasting new product sales, it is essential that we acknowledge and briefly discuss the repeat purchase component as well. A natural extension of this paper would be to conduct a similar 'M-competition' among repeat purchase models, or perhaps among combined trial-repeat models. The thought of such an analytical exercise raises additional questions. For instance, some managers may favour a fully integrated trial-repeat model (such as STEAM), while others may prefer a more 'modular' combination of different, potentially unrelated, trial and repeat components. Is one of these approaches better at forecasting than the other? Some difficult tradeoffs must be made between the flexibility of a modular system versus the statistical benefits that might arise if consumers' true, underlying, trial and repeat tendencies are tightly linked and not easily made separable.

Given the number of different model specifications identified in this paper, as well as the additional variety of repeat models available in the literature (i.e. those associated with the trial models presented in this paper), it is clear that a complete analysis and comparison of new product sales models would be a sizeable task. While our coverage of trial only covers a subset of the issues at stake, our organizing framework (Table I) can help guide the classification of sales

models and suggest the key characteristics to examine in a complete empirical analysis. As suggested in the previous section, we believe that many or all of our principal findings would continue to hold true, but it will be very interesting to see how the relative importance of factors such as heterogeneity, estimation method, and the presence or absence of a 'never triers' parameter change as we move from trial to repeat purchasing.

Future Outlook

Viewed from a broad perspective, the world of new product trial models has changed very little over the past 20–30 years. Perhaps this lack of attention is due to a belief on the part of academic modellers and practitioners that the various models in the literature are all fairly robust and that there is little difference between them in terms of performance. Our research has shown that any such belief is clearly wrong—the choice of model *does* matter. Academics should view the above-mentioned set of unresolved issues as good motivations for future research, and practitioners should begin to see the benefits of choosing models and methods more carefully (as well as the opportunity costs of not doing so).

A number of recent developments make the execution and implementation of these forecasting models even more feasible and attractive than ever before. First and foremost are computational reasons. All the model estimations performed in this paper were done using the standard 'Solver' utility that is built into Microsoft Excel—the authors used no 'academic' or custom software whatsoever. Second, each of these models require only seconds to run within Excel. (In fairness, however, the process of running and checking all 874 models for this paper was rather time-consuming.) Thus, today's managers have more computational power at their disposal than any academic had at the time that these models were first developed.

A related issue has been the growing desire to create and use databases across numerous new product launches. As the concept of 'data mining' continues to spread in popularity, managers are beginning to be more systematic in their approach to new product introduction decisions. This results in better record keeping, more uniformity across launches, and more motivation to track results. In our view, the key to success in this area is to build databases of model parameters and forecasts, rather than emphasizing raw data, *per se*. 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 trial/repeat models.

Finally, recent years have also seen significant improvements in panel data collection techniques, which provide further motivation to develop new and richer forecasting models. The old handwritten purchasing diaries were eclipsed by electronic laser scanners a generation ago, but newer developments make data collection easier and more representative than ever before. One prominent example is the notion of a multi-outlet purchasing panel, which provides household-level data for purchases from convenience stores, pharmacies, mass merchandisers, and warehouse clubs, in addition to supermarkets. As manufacturers rely more and more on staggered new product rollouts across different store types and geographic areas (Crawford, 1997), it is all the more essential to be able to develop and implement reliable forecasting procedures to allow for mid-stream adjustments to the marketing/production plans, as well as other corrective actions.

ACKNOWLEDGEMENTS

The authors thank Information Resources, Inc. for its assistance and the provision of data, and Rob Stevens, Lisa Cain, Len Lodish, Scott Armstrong, John Roberts, and Robert Fildes for

their input at various stages of this research. The first author acknowledges the support of the London Business School Research & Materials Development Fund and the LBS Centre for Marketing.

REFERENCES

Aaker, D. A., 'The new-trier stochastic model of brand choice', *Management Science*, **17** (1971), B435–B450.
Anscombe, F. J., 'Estimating a mixed-exponential response law', *Journal of the American Statistical Association*, **56** (1961), 493–502.
Armstrong, J. S. and Collopy, F., 'Error measures for generalizing about forecasting methods: empirical comparisons', *International Journal of Forecasting*, **8** (1992), 69–80.
Bass, F. M., 'A new product growth model for consumer durables', *Management Science*, **15** (1969), 215–27.
Bass, F. M., Krishnan, T. V. and Jain, D. C., 'Why the Bass model fits without decision variables', *Marketing Science*, **13** (1994), 203–23.
Blattberg, R. and Golanty, J., 'TRACKER: An early test-market forecasting and diagnostic model for new-product planning', *Journal of Marketing Research*, **15** (1978), 192–202.
Burger, P. C., 'Developing forecasting models for new product introductions', in King, R. L. (ed.), *Marketing and the New Science of Planning*, Chicago, IL: American Marketing Association, 1968, 112–18.
Clancy, K. J., Shulman, R. S. and Wolf, M., *Simulated Test Marketing: Technology for Launching Successful New Products*, New York: Lexington Books, 1994.
Crawford, C. M., *New Products Management*, 5th edn, Burr Ridge, IL: Irwin, 1997.
Curry, D. J., *The New Marketing Research Systems*, New York: John Wiley, 1993.
Ehrenberg, A. S. C., 'The pattern of consumer purchases', *Applied Statistics*, **8** (1959), 26–41.
Eskin, G. J., 'Dynamic forecasts of new product demand using a depth of repeat model', *Journal of Marketing Research*, **10** (1973), 115–29.
Eskin, G. J., 'Causal structures in dynamic trial-repeat forecasting models', *1974 Combined Proceedings, Series No. 36*, Chicago, IL: American Marketing Association, 1974, 198–201.
Eskin, G. J. and Malec, J., 'A model for estimating sales potential prior to the test market', *Proceeding 1976 Fall Educators' Conference, Series No. 39*, Chicago, IL: American Marketing Association, 1976, 230–3.
Fildes, R., 'The evaluation of extrapolative forecasting methods', *International Journal of Forecasting*, **8** (1992), 81–98.
Fourt, L. A. and Woodlock, J. W., 'Early prediction of market success for new grocery products', *Journal of Marketing*, **25** (1960), 31–8.
Gatignon, H. and Robertson, T. S., 'A propositional inventory for new diffusion research', *Journal of Consumer Research*, **11** (1985), 849–67.
Greene, J. D., 'Projecting test market "trial-repeat" of a new brand in time', *1974 Combined Proceedings, Series No. 36*, Chicago, IL: American Marketing Association, 1974, 419–22.
Herniter, J., 'A probabilistic market model of purchase timing and brand selection', *Management Science*, **18**, Part II (1971), P102–P113.
Johnson, N. L., Kotz, S. and Balakrishnan, N., *Continuous Univariate Distributions, Volume 1*, 2nd edn, New York: John Wiley, 1994.
Jones, J. M. and Mason, C. H., 'The role of distribution in the diffusion of new durable consumer products', Report No. 90-110, Cambridge, MA: Marketing Science Institute, 1990.
Kalwani, M. U. and Silk, A. J., 'Structure of repeat buying for new packaged goods', *Journal of Marketing Research*, **17** (1980), 316–22.
Lawrence, R. J., 'The penetration path', *European Research*, **7** (1979), 98–108.
Lawrence, R. J., 'A lognormal theory of purchase incidence', *European Research*, **10** (1982), 154–63. [See *European Research*, **11** (1983), p. 9 for corrections.]
Lawrence, R. J., 'The first purchase: models of innovation', *Marketing Intelligence and Planning*, **3** (1985), 57–72.

Mahajan, V., Mason, C. H. and Srinivasan, V., 'An evaluation of estimation procedures for new product diffusion models', in Mahajan, V. and Wind, Y. (eds), *Innovation Diffusion Models of New Product Acceptance*, Cambridge, MA: Ballinger, 1986, 203–32.

Mahajan, V., Muller, E. and Bass, F. M., 'New product diffusion models in marketing: a review and directions for research', *Journal of Marketing*, **54** (1990), 1–26.

Mahajan, V., Muller, E. and Sharma, S., 'An empirical comparison of awareness forecasting models of new product introduction', *Marketing Science*, **3** (1984), 179–97.

Mahajan, V. and Wind, Y., 'New product forecasting models: directions for research and implementation', *International Journal of Forecasting*, **4** (1988), 341–58.

Makridakis, S., 'Accuracy measures: theoretical and practical concerns', *International Journal of Forecasting*, **9** (1993), 527–29.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., 'The accuracy of extrapolative (time series) methods: results of a forecasting competition', *Journal of Forecasting*, **1** (1982), 111–53.

Massy, W. F., 'Stochastic models for monitoring new-product introduction', in Bass, F. M., King, C. W. and Pessemier, E. A. (eds), *Applications of the Sciences in Marketing Management*, New York: John Wiley, 1968, 85–111.

Massy, W. F., 'Forecasting the demand for new convenience products', *Journal of Marketing Research*, **6** (1969), 405–12.

Massy, W. F., Montgomery, D. B. and Morrison, D. G., *Stochastic Models of Buying Behavior*, Cambridge, MA: The MIT Press, 1970.

Meade, N., 'The use of growth curves in forecasting market development—a review and appraisal', *Journal of Forecasting*, **3** (1984), 429–51.

Morrison, D. G. and Schmittlein, D. C., 'Generalizing the NBD model for customer purchases: what are the implications and is it worth the effort?' *Journal of Business and Economic Statistics*, **6** (1988), 145–59.

Nakanishi, M., 'Advertising and promotion effects on consumer response to new products', *Journal of Marketing Research*, **10** (1973), 242–9.

Narasimhan, C. and Sen, S. K., 'New product models for test market data', *Journal of Marketing*, **47** (1983), 11–24.

Parfitt, J. H. and Collins, B. J. K., 'Use of consumer panels for brand-share prediction', *Journal of Marketing Research*, **5** (1968), 131–45.

Parker, P. M., 'Aggregate diffusion forecasting models in marketing: a critical review', *International Journal of Forecasting*, **10** (1994), 353–80.

Parker, P. and Gatignon, H., 'Specifying competitive effects in diffusion models: an empirical analysis', *International Journal of Research in Marketing*, **11** (1994), 17–39.

Pringle, L. G., Wilson, R. D. and Brody, E. I., 'NEWS: a decision-oriented model for new product analysis and forecasting', *Marketing Science*, **1** (1982), 1–29.

Rao, S., 'An empirical comparison of sales forecasting models', *Journal of Product Innovation Management*, **2** (1985), 232–42.

Srinivasan, V. and Mason, C. H., 'Nonlinear least squares estimation of new product innovation models', *Marketing Science*, **5** (1986), 169–78.

Urban, G. L., 'Pretest market forecasting', in Eliashberg, J. and Lilien, G. L. (eds), *Handbooks in Operations Research and Management Science, Volume 5: Marketing*, Amsterdam: Elsevier Science Publishers BV, 1993, 315–48.

Urban, G. L., 'SPRINTER Mod III: a model for the analysis of new frequently purchased consumer products', *Operations Research*, **18** (1970), 805–54.

Urban, G. L. and Katz, G. M., 'Pre-test market models: validation and managerial implications', *Journal of Marketing Research*, **20** (1983), 221–34.

Van den Bulte, C. and Lilien, G. L., 'Bias and systematic change in the parameter estimates of macro-level diffusion models', *Marketing Science*, **16** (1997), 338–53.

Wilson, R. D. and Pringle, L. G., 'Modeling new-product introductions: a comparison of NEWS, SPRINTER, and TRACKER', in Srivastava, R. K. and Shocker, A. D. (eds), *Analytic Approaches to Product and Marketing Planning: The Second Conference*, Report No. 82-109, Cambridge, MA: Marketing Science Institute, 1982, 297–311.

Authors' biographies:
Bruce G. S. Hardie is an Assistant Professor of Marketing at the London Business School. His research interests centre on probability modelling, choice modelling, and market share modelling, especially as applied to substantive marketing problems in new product forecasting, product line management, and sales promotions. His research has appeared in *Marketing Science*, *Journal of Marketing Research*, *Marketing Letters*, and other journals.

Peter S. Fader is an Associate Professor of Marketing at the Wharton School of the University of Pennsylvania. His research focuses on uses of data generated by new information technology, such as supermarket scanners, to understand consumer preferences and to assist companies in fine-tuning their marketing tactics and strategies. He has been published in many fine journals, and was recently named by *People* magazine as one of the 25 sexiest men in the world.

Michael Wisniewski is a PhD student in Marketing at the London Business School. His research interests are in mathematical modelling applied to both strategic and tactical problems within marketing.

Authors' addresses:
Bruce G. S. Hardie and **Michael Wisniewski**, London Business School, Sussex Place, Regent's Park, London NW1 4SA, UK.

Peter S. Fader, Marketing Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6371, USA.

