Forecasting New Product Trial with Analogous Series

Malcolm J. Wright, Massey University and Ehrenberg-Bass Institute, University of South Australia and

Philip Stern, Exeter University and Ehrenberg-Bass Institute, University of South Australia

Submission: October 2014

Revision: February 2015

Acceptance: January 2015

The authors thank Kesten Green, Scott Armstrong, Bruce Hardie, Stavros Kalafatis, Nigel Meade, Viv Shaw, Towhidul Islam, Jim Wiley and the anonymous reviewers for comments on earlier versions of this paper. Send correspondence to Malcolm J. Wright (corresponding author), College of Business, Massey University Manawatu, Private Bag 11 222, Palmerston North 4442, New Zealand. Telephone: +64 (06) 356 9099 ext. 83974, email: M.J.Wright@massey.ac.nz; Philip Stern, University of Exeter Business School, Rennes Drive, Exeter EX4 4PU, United Kingdom Telephone: +44 (0) 1392 725971 email: p.stern@exeter.ac.uk.

.

Forecasting New Product Trial with Analogous Series

Abstract

This study develops a simple method for forecasting consumer trial for national product launches. The number of consumers who try a brand in its first year on the market is accurately predicted from the number trying the brand in the first thirteen weeks following launch. No information about the specific category or marketing activities is required– just a simple multiplier computed from analogous series in other markets. These analogues provide an empirical generalization that can be easily applied by practicing managers to track and forecast the success of new brand launches. When subject to an out-of-sample test involving 34 fresh data sets, the analogues demonstrated 43 percent reduction in Mean Absolute Percentage Error compared to the most accurate marketing science model.

Keywords: New Products, Consumer Trial, Penetration Growth, Analogous Series, Empirical Generalization Pharmaceuticals, Packaged Goods.

1. Introduction

Accurate forecasts of new product performance have many uses: they help to decide whether to launch the new product; they guide production plans and optimal stock levels; they encourage sales and retailer support; and they provide benchmarks for sales targets over the first year. Although there are many complex marketing science models available for making new product forecasts, these complex models are rarely (if ever) used by managers, and when forecasting is undertaken its performance remains notoriously poor. For example, Khan's (2002) survey of marketing managers asked them to report the level of accuracy achieved for their new product forecasts; he found that self-reported forecast accuracy averaged just 58 percent, dropping to 47 percent for new category entrants and 40 percent for products that were new to the world.

The present work addresses one part of this problem, by developing forecasts and benchmarks for growth in consumer trial over the first year of sales. Consumer trial is the proportion of the target market who purchase once or more in a given period after product launch. Consumer trial is a simple measure and one of the most commonly used metrics for evaluating new product launch as it is easy to gather via syndicated panel data or ad-hoc tracking surveys (e.g. using the question "have you ever bought/tried this product?"). Logically, consumer trial is a hurdle which must be cleared before repeat purchase and long-term sales can be established.

Strictly speaking, the first purchase is not the same as the sociological concept of a trial. Trialability is an important characteristic of innovations in the diffusion literature, and may be promoted through, for example, a test drive for a car, product sampling or a first purchase of a frequently consumed product. For durables such as cars, first purchases are adoption rather than trial, and even if followed by replacement or disenchantment discontinuance the purchase was nonetheless adoption and not just trial; see Rogers (1995) for further discussion of these concepts. For frequently purchased products, there is some argument about whether the first purchase represents trial, immediate adoption, or simply a rate of purchase that may be revised up or down depending upon satisfaction with the product experience; see Fader et al. (2003) and East, Wright, and Vanhuele (2013) for a discussion of this issue and the associated literature. In a similar vein, Meade and Islam (2001) distinguish the adoption process (e.g. for durables) from a consumption process (e.g., for packaged goods) in which adoption choices are revocable. Despite these complications, common usage applies the term *trial* to the cumulative penetration metrics obtained for new products from consumer surveys for consumption processes. The present work follows common usage, with *consumer trial* referring to those who have bought a newly launched frequently purchased product at least once.

Consumer trial has long been forecast using marketing science models, although there are few systematic evaluations of the performance of these models. Meade and Islam (2001) assess adoption models in which choices are not easily revocable. Hardie, Fader and Wisniewski (1998) – hereafter HFW – evaluate consumption models for frequently purchased products in BehaviorScan controlled test markets (see also Fader and Hardie 2001 and Fader, Hardie and Zeithammer 2003).

More recent work extends these modeling approaches by introducing additional complexity. Du and Kamakura (2011) apply an individual-level trial hazard model to evaluate the role of contagion in the diffusion of new product trial. Liutec, Du, and Blair (2012) develop Du and Kamakura's model (2011) using BehaviorScan data; they describe their approach as follows:

> The actual forecasting procedure has several steps. First, estimates for the historical set of products are obtained. Second, these estimates are used as priors for estimation of

> the “forecasting model”, which is calibrated on data for both the focal product (first weeks available post launch, and incorporating more observations as they become available) and the historical adoption data. Third, posterior estimates for all (focal) product-level and individual level parameters are computed (e.g., Train 2003). Finally, the posterior estimates obtained above are used to simulate out-of-sample purchases for the focal product. More specifically, many out-of-sample “purchase paths” are simulated for each consumer, then averaged across the number of simulations and summed across consumers. (Liutec, Du. and Blair, (2012, p. 17)

The complexity of Liutec et al’s approach may prevent widespread utilization by marketing managers. Further, much of the evaluation of the consumer trial models takes place in BehaviorScan controlled test markets, and it is unclear how well these models will generalize to the different context of national product launches.

While the BehaviorScan markets used by HFW and Liutec at al. (2012) possess many attractive features, such as high levels of marketing control, immediate distribution, and provision of complete information on purchases, they are restricted to small test cities with relatively homogenous populations. BehaviorScan market tests are expensive, are only available in a few countries, and are restricted to grocery products. They are not necessarily representative of broader domestic or international markets. In contrast, most brand managers rely on panel tracking or survey data to assess the performance of new products. The trial curves observed in test markets may not generalize to national launches due to the uncontrolled environment and varied populations involved. These varied populations may adopt simultaneously, as with test markets, or serially, and this may vary between countries (e.g. see Meade and Islam 2001).

Therefore, the first objective of the present work is to extend the analysis of forecasting performance of the marketing science models of consumer trial from test markets to national launches. This provides baseline results against which a simplified forecasting technique can be compared. Simplified forecasting techniques are desirable as even the best marketing science model may be limited in its practical application. Khan (2002) finds the more popular statistical forecasting techniques for new products are trend line and moving average analysis. Techniques such as diffusion, ARIMA or regression are rarely utilized, and one reason for this could be the impact of a 'bias' in forecasting Gigerenzer and Brighton, (2009); Brighton and Gigerenzer (this issue). Overall, Khan (2002) finds a preference for qualitative techniques.

Analogous series is a simpler quantitative technique that can be applied to forecast consumer trial. Duncan, Gorr and Szczpula (2001) argue that pooling of analogous series improves forecasting accuracy when time series are highly volatile or when they have outlier data points. As consumer trial frequently shows early fluctuations in the data (volatility) analogous series are an appropriate method of analysis. If successful, they may provide an empirical generalization that is easily applied by managers, helping to resolve the forecasting performance issues identified by Khan (2002). That is, analogous series would provide managers with a simple and successful quantitative tool for forecasting consumer trial of new products.

Therefore, the second objective is to apply the simpler technique of analogous series to forecasting consumer trial, and to compare the results with those of the marketing science models. Analogous series comprise

> data that are expected to be related and are conceptually similar. Such series are expected to be affected by similar factors (Armstrong, 2001, p. 764).

The final objective is to evaluate the results for analogous series using out-of-sample tests with fresh data sets. This gives more confidence in the generalizability of the findings.

The remainder of the paper is structured as follows: Section 2 introduces the data used in the initial studies. Section 3 extends the baseline marketing science models from 19 controlled test market product launches to the context of 12 national product launches. Section 4 develops and applies the analogous series to these 31 examples of controlled test markets and national product launches. Section 5 applies the analogues for national product launches to a further 34 out-of-sample test data sets. The paper concludes with a discussion and summary.

2. Data

The data for the baseline comparison of marketing science models are cumulative trial panel data sets for 12 new product launches from three countries. Commercial confidentiality prevents full description, but a summary appears in Table 1. The data are a mix of really new products, line extensions and simple variants. The New Zealand data are provided courtesy of AC Nielsen Homescan. The USA data are provided by a marketing science consultant. The data on UK drug prescribing come from Jigsaw, a commercial panel of UK General Practitioners operated by Synovate. The four drugs analyzed are Didronel used to treat osteoporosis, and three new drugs used to treat hypertension: Cozaar was the first to market followed by Diovan (a me-too) two years later and Amias (another me-too) launched a year after Diovan.

Table 1 here**.**

The range of product categories studied is deliberately diverse to provide a strong test of the analogous series approach. As well as geographical differences there are varying degrees of innovativeness, and even general practitioners' prescriptions of new drugs. Although qualitatively very different, patterns of prescribing behavior for existing drugs are known to be similar to patterns of purchase behavior for packaged goods (Stern and Ehrenberg, 1995) and so are suitable for this analysis.

The results for these data are compared with the results for 19 BehaviorScan cumulative trial data sets provided online at brucehardie.com. Sample sizes for the BehaviorScan data sets range from 566 to 2946 and average 1146. No details are provided about the individual brands.

The performance of the analogous series prompts an examination of further data as a check on the initial results. This uses 34 fresh test data sets, consisting of cumulative trial panel data for GP's prescriptions of new drugs. Sample sizes range from 272 to 460 and average 373. The sample sizes are slightly smaller but of the same magnitude as those used by HFW. These data include new products from a variety of categories ranging from antidepressants, to drugs for lowering cholesterol, treatments for erectile dysfunction and rheumatoid arthritis, and widely differing degrees of innovativeness and success.

3. Method and Results – Baseline Models

The work of HFW guided the selection of the baseline marketing science models. Many marketing science models of consumer trial include the exponential, exponential gamma, Weibull-gamma, lognormal-lognormal, double exponential and Bass models. In some cases, these models are further complicated by the inclusion of a never-triers parameter or a stretch parameter. HFW compare the forecasting performance of eight such models using 52 weeks of data for each of 19 BehaviorScan data sets. They used both 13-week and 26-week calibration periods and examine forecasting performance for the remaining weeks using Forecast MAPE (mean absolute percentage error), to show the evolving performance of each model as well as the difference between forecast and actual trial at fifty-two weeks (52 Week APE, or absolute percentage error). They find the four simpler exponential trial growth models provide more accurate forecasts than the four more complex models (Weibull with never triers, Lognormal-Lognormal, Double Exponential and Bass). In subsequent work they abandon these complex models and also a fifth model, the gamma never-triers with stretch (Fader and Hardie 2001) due to the logical inconsistency of the formulation which can lead to trial proportions exceeding unity (Hardie 2006 Personal Communication).

The present work therefore restricts use of baseline marketing science models to the three that survived HFW's winnowing. These are, in order of decreasing simplicity; an exponential model with a segment of never-triers, the exponential gamma model (which is the purchase-timing analogue of familiar Negative Binomial Distribution model (Ehrenberg, 1959)), and the exponential-gamma model with a segment of never-triers.

3.1 Model Specification

HFW note their exponential with never-triers model arises from the well-known approach of Fourt and Woodlock (1960). They propose a continuous-time revision of this model to incorporate a segment of never-triers, as follows:

$$P\ (t) = p\ (1 – e^{–\lambda t}) \qquad (1)$$

where P(t) = cumulative trial of the new product at time t,

p = proportion that will ever try the new product,

λ = probability of trial, given that trial has not yet occurred, for those that will ever try the new product,

t = time period since launch.

Consumer heterogeneity is added to this model by allowing the purchase rate to vary across the population. This is achieved using a gamma distribution with scale parameter α and shape parameter r, leading to the following exponential-gamma model:

$$P\ (t) = 1 – (\alpha / (\alpha + t))^{r} \qquad (2)$$

Adding back the allowance for never-triers yields the third model, exponential-gamma with never triers:

$$P\ (t) = p\ [1 – (\alpha / (\alpha + t))^{r}] \qquad (3)$$

3.2 Estimation

Estimation is through non-linear least squares (NLLS) using cumulative trial. This involves minimizing the sum of the squared errors between estimated and actual cumulative trial using the iterative procedures available in standard commercial software packages. HFW find this approach produces better results than NLLS using non-cumulative trial, and similar results to maximum likelihood estimation. Although maximum likelihood estimation is theoretically superior, the empirical results are only marginally better (Fader and Hardie 2001, p625). NLLS can be computed using the Solver function in Excel, making both replication and practical application of this approach easier. Following HFW, model estimation is based on 13-week and 26-week calibration periods.

The USA data sets have some missing data. These data are aggregated into four-weekly intervals, so values for weeks 13 and 26 are not available. They are interpolated using a simple linear model. These imputations are used for model fitting only, and not to assess forecasting performance. Also, one New Zealand product launch has only 47 weeks of data with a maximum 183 cumulative trialists. The interpolated value for week 52 is imputed as 189 trialists; this value yields identical 52-week percentage errors to the other 11 data sets, and so does not bias the estimate of 52-week forecast performance.

3.3 Results

For the 13-week calibration period, Table 2 shows forecast period mean absolute percentage error (Forecast MAPE, also called Tracking MAPE) and 52-week forecast absolute percentage error (52w_APE) for the three models. The Forecast MAPE shows how closely the estimated cumulative trial curves track empirical cumulative trial, while the 52-week APE assesses the accuracy of the forecast of year-end cumulative trial. Each number represents the average across all twelve data sets, based on the 13-week calibration period. The equivalent results obtained by HFW from their 19 BehaviorScan test market data sets are included for comparison. Smaller numbers indicate lower forecast error.

Table 2 here.

Comparing the three rows of alternative models, the differences in relative performance within each set of data are small. In each case, any one model performs about as well as any other, although the simpler Exp.-Gamma model performs best overall. However, comparing the column means *between* the two sets of data, the Forecast MAPE for national launches is almost double that found using the BehaviorScan data (28 percent versus 16 percent). This is likely due to the greater variety among the national launches, including more differences between customers, more random disturbances and greater marketing variations when moving from BehaviorScan test markets to panel tracking data. Despite this, the 52-week APE values are reasonably close across both sets of data, albeit slightly smaller for the test markets.

Table 3 provides an identical analysis for the 26-week calibration period. As expected the longer calibration period results in improved forecasting performance. The results for national launches are now comparable with those found by HFW, and in fact are slightly better. This confirms that the original pattern of results found by HFW can be extended to national product launches, with the proviso that a 26-week calibration period is necessary to achieve comparable accuracy to BehaviorScan test markets.

Table 3 here.

These results are consistent with several of the forecasting principles reported by Fader and Hardie (2001): that is simpler models perform well; that including consumer heterogeneity (variation in purchase rates) improves forecasts; and that a never-triers parameter does not improve Forecast MAPE.

Despite the different patterns of trial growth, the marketing science models used in test markets can be extended to national panel data and this provides a baseline against which the performance of analogous series can be compared for both national product launches and BehaviorScan test markets.

4. Method and Results – Analogous Series

Implementing analogous series first requires that all cumulative trial curves are made directly comparable. This involves converting trial figures into proportions then, for each data set, calculating the ratio of Week <x> cumulative trial to 52-week cumulative trial, which normalize each period's trial to year-end values. For forecasting, the average of these values for any particular period *t*, is used to estimate the proportion of year-end cumulative trial achieved by period *t*.

Formally,

$$P_j\,(t) \quad = \quad (1 / n)\ \sum_{j = 1 \text{ to } n} \left(p_{jt} / p_{j52} \right) \qquad (4)$$

Where $P_j\,(t)$ = average proportion of 52-week cumulative trial achieved by time t, for the group of data sets j, where j varies from 1 to n

p_{jt} = cumulative trial for new product j at time t

p_{j52} = cumulative trial for new product j at Week 52

Equation (4) uses all the data available to identify the analogous empirical pattern. While this is the best estimate of the analogous series for future applications, this is not appropriate for assessment of forecasting performance; by including all the data sets, Equation (4) includes the values to be forecast. Assessing the accuracy of analogous series requires a modification to omit the data set under consideration (j=i), as follows.

$$P_i\,(t) \quad = \quad (1 / (n\text{-}1))\ \sum_{j = 1 \text{ to } n,\ j \neq i} (p_{jt} / p_{j52}) \qquad (5)$$

This gives, for each data set, the analogous pattern present in all the other data sets. This is used to forecast the focal data set as shown Equation (6). This specification allows measures of forecast error to be calculated in exactly the same way as for the exponential trial growth models in Equations (1) to (3).

$$P_{i52} \quad = \quad p_{it} \; / \; P_j \,(t) \qquad (6)$$

In BehaviorScan test markets cumulative trial grows more quickly than in national product launches, probably due to their homogenous populations, immediate distribution and tightly managed marketing programs. This is part of the purpose of BehaviorScan test markets; as by achieving trial more quickly they allow a faster assessment of the test brand. Thus separate sets of analogues are calculated for national product launches and BehaviorScan test markets.

Table 4 shows the results. The second and third columns report the proportions of year-end trial, which are the average ratios across all data sets using equation (4); in other words, the analogous series. The next two columns show the 52w_APE values for the estimates derived from equation (5) and (6).

Table 4 here

The forecasting errors (52w_APE) are highest soon after launch when the least data is available. Forecast errors from national launches are higher than those from BehaviorScan test markets initially, but become lower once 13 weeks of data are available. Consistent with another principle that Fader and Hardie (2001) propose, lengthening the calibration period eventually ceases to have much impact on accuracy – in this case, at around 26 weeks.

So far, this pattern of results is qualitatively similar to that found for the marketing science models. In order to assess how analogous series perform, the 52w_APE values in Table 4 are compared with the 52w_APE values in Tables 2 and 3.

For national product launches with a 13-week calibration period, the analogous series value of 13 is much lower than the values of 27, 29 and 30 for the marketing science models. For a 26-week calibration period, the analogous series value of 6 is much lower than the values of 9, 9, and 10 for the marketing science models. (US national launches do not include data for the 26-week period; however, they do include data for the 24-week period, and the analogous series for 24-weeks has a 51w_APE of 10.)

For BehaviorScan test markets with a 13-week calibration period, the analogous series value of 23 is similar to or lower than the values of 23, 24 and 24 for the marketing science models. For 26-week calibration period the analogous value of 11, is lower than the values of 12, 12 and 18 for the marketing science models

Thus, the analogous series provide comparable or better estimates than the marketing science models. Analogous series dominate the three more complicated models. In the case of the simplest exponential-gamma model, the analogous series has lower forecast error for three of the comparisons, and equivalent forecast error for the fourth.

5. Method and Results: Additional Data sets

The initial results are surprising and so should be checked using challenging tests against competing approaches. This is accomplished with the use of 34 additional pharmaceutical data sets, obtained after the initial analysis was completed, to determine how well the analogous series for the original 12 national product launches predicts for genuinely new data.

Figure 1 presents the average of the normalized cumulative trial proportions for all three data sets – the 19 HFW BehaviorScan test market data sets, the 12 national product launch data sets, and the 34 extra pharmaceutical test data sets. In other words, Figure 1 presents the two analogous series from Table 4, and a new analogous series calculated from the 34 additional pharmaceutical test data sets.

Figure 1 here.

The pattern of growth in normalized cumulative trial proportions for test markets differs from the other series; this is as expected, due to the controlled conditions present in the test market. In contrast, the analogous series for national launches and extra pharmaceutical data sets are similar. This similarity implies the analogous series approach will fit the extra pharmaceutical data sets on average; however, this does necessitate accurate forecasting performance, as individual data sets vary considerably around the series average, as appearing in Figure 2.

Figure 2 here

The competing approaches against which the analogous series is assessed are a naïve straight-line growth in cumulative trial, and the exponential-gamma marketing science model. As before, forecasts are made at 13 and 26 weeks, and evaluated using Forecast MAPE and 52w_APE, together with a new measure of 52w_MPE, defined as the mean percentage error, or bias, of the 52 week forecast. A positive value of 52w_MPE indicates the actual trial is higher than the forecast trial.

Table 5 shows the results, including a summary of the performance of the exponential gamma model previously reported in Tables 2 and 3. These results are not directly comparable with the new analysis, but do provide a reminder of the magnitude of typical forecast errors for the best performing marketing science model.

Table 5 here**.**

The naïve comparator of a straight line provides a surprising challenge, with forecast (tracking) MAPE within the range of values typically found for the exponential-gamma model. However, the exponential-gamma model does still outperform the naïve comparator and substantially so on 52w_APE and 52w_MPE.

More generally, the fit of the exponential-gamma model is almost as accurate as for the other 31 data sets examined. While forecast errors are a few points higher, this is not surprising given the smaller sample sizes in the 34 test data sets (averaging 373 compared to 1146 for BehaviorScan and 2039 for the national product launches).

Nonetheless, the analogous series out-performs both other methods. For the 13-week calibration period the analogous series dominate the exponential-gamma method over all three measures of forecast error with a 43 percent reduction in Forecast MAPE, an 8 percent reduction in 52w_APE, and a lower level of bias as shown by 52w_MPE. For the 26-week calibration period, the analogous series out-perform the exponential gamma method on two measures of forecast error with a 38 percent reduction in Forecast MAPE and a 14 percent reduction in 52w_APE, and closely follow on the third measure of 52w_MPE. Further, the value of 2 percent for 52w_MPE indicates minimal bias arising from the analogous series approach in the 26-week calibration period. Again the simplest model – in this case the analogous series - works best, consistent with forecasting principles proposed by Duncan, Gorr and Szczpula (2001), Fader and Hardie (2001), and Meade and Islam (2001).

Should the analogous series in Table 4 include values from the 34 test data sets? The appropriateness of doing so is unclear, and the differences are in any event trivial – no more than .01 for the two estimation periods, and seldom more than this for any subsequent period. Therefore, rather than pool the test data set with the estimation data set, the present work leaves the analogous series unchanged from those in Table 4.

These analogues provide: (i) a benchmark for cumulative trial growth; and (ii) projection of year-end results. Consider benchmarking: if cumulative trial in a national launch is forecast to be 50 percent at year-end, and only 20 percent has been achieved by week 16, is this high, low or just right? Multiplying the year-end forecast (.50) by the analogy from Table 4 (.35 from week 16) yields forecast cumulative trial of 17.5 percent. Therefore, a rate of 20 percent is on target, or even slightly high. Conversely, if cumulative trial has reached 15 percent by week 26, the analogue can be used to forecast the end of year trial rate. The analogue for week 26 is .62 of year-end trial. So the forecast for year-end trial is simply .15/.62 or 24 percent.

As an illustration and further test of the method, the analogues in Table 4 can be applied to data from Singh, Scriven, Clemente, Lomax, and Wright (2012). They report in passing the average quarterly cumulative penetration for 47 brand extensions in UK packaged goods categories. In this case the quarters are 3 x 4 week periods, or 12 rather than 13 weeks. The appropriate analogues from Table 4 are therefore: the 12-week value for the first quarter (Q1); and, the 48-week value for the fourth quarter (Q4). Using these analogues, Q4 cumulative penetration is forecast to be .95/.25 or 3.82 times the Q1 value. The actual Q1 value in Singh et al. (2012) is 1.4 percent so the forecast Q4 value is 5.35 percent (3.82 * 1.4 percent). This is within 6 percent of the observed Q4 value of 5.7 percent. This provides additional support for the use of the analogues in Table 4.

Anecdotally, practitioners know they need to wait a few months for distribution to settle down before interpreting market performance data. Therefore, forecasting year-end cumulative trial from first quarter results, after distribution is established, is likely a very common problem. Table 4 indicates that 25 percent of year-end trial will be achieved after 12 weeks, and 31 percent after 13 weeks. To aid memorability, and in line with the principle of conservatism in forecasting (Armstrong, Green and Graefe, this issue), the second figure could reasonably be rounded up to one third. So managers wishing to know year-end cumulative trial could simply multiply the cumulative trial achieved at week 12 by four, or that achieved at week 13 by three. Given the rapid growth of consumer trial between 12 and 13 weeks, a 13-week time period should be used in preference to a 12-week time period, if possible.

6. Discussion

This work first investigates whether exponential trial growth models can be generalized beyond controlled test markets. That is, given the lesser control and greater variability of national markets, will these models still apply? This research yields a clear answer – the *shape* of cumulative trial curves in national markets is different, but the exponential trial growth models do still apply.

Next, this research investigates whether analogous series can provide forecasts as accurate as the marketing science models. Again, the answer is clear - they do, in both BehaviorScan test markets and national product launches, and this enables a simple table of ratios to be used in place of the marketing science models. Aside from the practical applications of this result, this confirms the worth of the analogous series approach (Armstrong, Green and Graefe, 201X, in this issue), for which empirical validation has been somewhat lacking (Armstrong 2001, p696, Principle 7.6).

Commented [P1]: Copy editor please insert correct year

These results rest on a firm empirical base. To assess the method, each series is forecast separately while being held out from the estimation sample, and this procedure is repeated in two differentiated contexts. Validation of the method therefore involves 31 hold out tests including a differentiated replication. Furthermore, the original analogues are used to derive forecasts for a number of completely fresh data sets. For an additional 34 test data sets, the original analogues give more accurate forecasts than both a naive comparator and the best performing marketing science model. This is surprising as, unlike the analogues, the marketing science model was re-estimated for each of these 34 test data sets. Finally we apply the original analogues to a further 47 data sets previously reported in the literature and find that the forecast is within 6% of the observed value.

These results fulfill Bass’ (1995) criteria for an empirical generalization, as a “pattern or regularity that repeats over different circumstances and can be described simply by mathematical, graphic or symbolic methods”. These results also meet Barwise’s (1995) criteria for a *good* empirical generalization, namely being based on repeated evidence and having scope, precision, parsimony, usefulness and a link with theory (i.e. stochastic models of consumer behavior).

This generalization does stand in contrast to Meade and Islam’s (2001, p. 583) view that, “Even for diffusion processes that are homogenous in the sense that they describe the same innovation in different geographical areas, there is evidence that no single model performs well in all cases." However, aside from confirming Meade and Islam's (2001, p. 587) simplicity principle, the results directly support their principle that, "unconditional forecasts based on a data-based estimate of a fixed saturation level form a difficult benchmark to beat."

The emergence of this empirical generalization also highlights one of the weaknesses of statistical modeling. Statistical modeling can be very good at identifying small effects and variations between markets; however, seeking best fits to individual series may miss large effects and patterns that are the *same* in many markets. The identification of these similar patterns is a strength of both the analogous series and empirical generalization approaches. Of course more sophisticated problems, such as early forecasting incorporating marketing covariates, or trial-repeat modeling of long-term total demand, will require more sophisticated modeling. The present work makes no claim to solve such problems. Nonetheless, Khan's (2002) results suggest that many practitioners do not undertake such sophisticated modeling, and for them the empirical generalization arising from the analogous series forecasting has practical application. Given Khan's (2002) findings on management forecasting practice, complex approaches of the type Du and Kamakura (2011) and Liutec et al. (2012) report may not find widespread application. Also, their approaches are only validated for consumer packaged goods in one geographical location. In contrast, the simpler approach reported in the present work predicts trial for new launches, in different geographical locations, and for highly varied types of product.

Ideally, those seeking to use this approach in new situations should develop their own analogues for these new situations, as follows: gather 6 to 12 previous time series of weekly cumulative trial data for new products; for each new product, divide each week by the cumulative trial achieved at the end of the first year; average the time series across products to yield a table of ratios similar to that in Table 4. As the national product launch data analyzed in this study is quite varied, individual applications are unlikely to yield very different results. Nonetheless, special circumstances sometimes apply–as with the different analogues for the controlled test-markets – so checking the results in each new context would be wise. The analogues may be further decomposed to match different situations; for example, first movers versus followers, or successes versus failures. This decomposition remains a matter for further research, for which the techniques described in Green and Armstrong (2007) will provide a useful starting point.

7. Summary

This research extends findings from BehaviorScan test markets to national product launches monitored by consumer panels, and compares traditional marketing science forecasting methods with the simpler approach of analogous series. As expected, longer calibration periods are required for comparable accuracy in the messier environments of a national product launch. Nonetheless, the broad pattern of results for marketing science models is consistent between the two types of markets. Also, consistent with Armstrong (2001), Fader and Hardie (2001) and Meade and Islam (2001), simpler models perform well, the inclusion of consumer heterogeneity improves forecasts, and a never-triers parameter does not improve Forecast MAPE. These results should extend managerial and academic confidence in the use of marketing science models from BehaviorScan test markets to panel data for national product launches. Yet the simpler approach of analogous series forecasting performs just as well as the more complicated statistical models, and dominates the best performing marketing science model in the case of a challenging test involving 34 further new product launch data sets.

Analogous series do currently offer less scope for the inclusion of more complex behaviors; such as the effect of marketing covariates, or the revision of purchase probabilities after product experience. However, the use of structured analogies (Green and Armstrong 2007) may enable the analogous series approach to be extended to these more complex behaviors, and may also provide help to guidelines for disaggregating series into successful and unsuccessful new product launches. These remain matters for future research.

Nonetheless, based on the present research, managers can use this approach for forecasting and benchmarking the simple cumulative trial metric widely used in new product launches. Managers may often simply wish to know, if they have achieved a certain level of trial after the first quarter, where will they be at the end of the year? Analogous series provide a simpler and better forecast than those available from marketing science models.

References

Armstrong, J.S, Green, K.C. and Graefe, A. (201X). Golden rule of forecasting: be conservative. Journal of Business Research, this issue, XXXX.

Commented [P2]: Copy editor please insert correct year, edition and pagination

Armstrong, J.S. (2001). Standards and Practices for Forecasting, in Armstrong, J. Scott (ed.). Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Press.

Bass, F.M. (1995). Empirical generalizations and marketing science: A personal view. Marketing Science 14(3): Part 2 of 2, G6-G18.

Barwise, P. (1995). Good empirical generalizations. Marketing Science 14(3): Part 2 of 2, G29-G35.

Brighton, H. and Gigerenzera, G. (201X) The bias bias. Journal of Business Research

Commented [P3]: Copy editor please insert correct year, edition and pagination

Du, Rex and Wagner Kamakura, (2011) Measuring Contagion in the Diffusion of Consumer Packaged Goods, Journal of Marketing Research, February, 28-47.

Duncan, George, T. Gorr, Wilpen, L., and Szczypula, Janusz (2001). Forecasting Analogous Time Series, in Armstrong, J. Scott (ed.) (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Press.

East, R., Wright, M. and Vanhuele, M (2013). Consumer Behaviour: Applications in Marketing. Sage: London.

Ehrenberg, A. S. C. (1959). The Pattern of Consumer Purchases. Applied Statistics 8(1): 26-41.

Ehrenberg, A.S.C. (1994). Theory or well-based results: Which comes first? In G. Laurent, G., Lilien, G.L. and Pras, B. (Eds.) Research Traditions in Marketing (pp 79-105, 1st ed.) Boston: Kluwer.

Fader, P. S., and Hardie, B. G. S. (2001). Forecasting Trial Sales of New Consumer Packaged Goods. In J. S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners (613-630). Norwell, Massachusetts: Kluwer.

Fader, P. S., Hardie, B. G. S., and Zeithammer, R. (2003). Forecasting new product trial in a controlled test market environment. Journal of Forecasting 22(5): 391-410.

Fourt, L. A., and Woodlock, J. W. (1960). Early Prediction of Market Success for New Grocery Products. Journal of Marketing 25(2): 31-38.

Gigerenzer, G. and Brighton, H. (2009) Topics Homo Heuristicus: Why Biased Minds MakeBetter Inferences Topics in Cognitive Science 1 107–143.

Green, K.C. and Armstrong, J.S. (2007). Structured analogies for forecasting. International Journal of Forecasting 23(3): 365-376.

Hardie, B. G. S., Fader, P. S., and Wisniewski, M. (1998). An empirical comparison of new product trial forecasting models. Journal of Forecasting 17(3-4): 209-229.

Khan, K.B. (2002). An exploratory Investigation of new product forecasting practices. Journal of Product Innovation Management 19: 133-143.

Lindsay, R.M., and Ehrenberg, A.S.C. (1993). The Design if Replicated Studies The American Statistician 47: 217-228.

Liutec, C., Du. R., and Blair, E. (2012) Investigating New Product Purchase Behavior: A multi-Product Individual-Level Model for New Product Sales, http://www.clsbe.lisboa.ucp.pt/resources/Documents/PROFESSORES/SEMINARIOS/2012%20Paper%20Carmen%20Liutec.pdf , accessed May 29th 2013)

Meade, N., and Islam, T. (2001). Forecasting the Diffusion of Innovations, Implications for Time-Series Extrapolation, in Armstrong, J. Scott (ed.) (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Press.

Rogers, E. (1995). Diffusion of Innovations, 4th Edition. The Free Press: New York.

Singh, J., Scriven, J., Clemente, M., Lomax, W., & Wright, M. (2012). New brand extensions: patterns of success and failure. Journal of Advertising Research, 52(2), 234-242.

Stern, P., & Ehrenberg, A. S. C. (1995). The market performance of pharmaceutical brands. Marketing and Research Today November: 285–292.

Table 1

Summary of Data Sets for National Launches

Behavior	New Product	n	Country	Innovativeness
Household Purchases	Biscuit Flavor	1719	New Zealand	Brand Variant
Household Purchases	Health Snack	1569	New Zealand	Category Entry
Household Purchases	Yellow Fat	2020	New Zealand	Line extension
GP Prescribing	Drug Molecule	322	UK	New to the UK
GP Prescribing	Drug Molecule	370	UK	Category Entry
GP Prescribing	Drug Molecule	441	UK	Category Entry
GP Prescribing	Drug Molecule	394	UK	Category Entry
Household Purchases	Yellow Fat	3000	USA	Line Extension
Household Purchases	Yellow Fat	3000	USA	Line Extension
Household Purchases	Yellow Fat	3000	USA	Category Entry
Household Purchases	Salad Dressing	4631	USA	Line Extension
Household Purchases	Mayonnaise	4000	USA	Line Extension

Table 2: Average Results for 13-week Calibration Period

Model	National Launches – 12 Panel Tracking Data sets		Test Markets (HFW) – 19 BehaviorScan Data sets	
	Forecast MAPE	52w_APE	Forecast MAPE	52w_APE
Exp.-Gamma	27	28	17	23
Exp.-Gamma w /NT	29	28	16	24
Exp. w/NT	30	31	16	24
Mean	***28***	***29***	***16***	***24***

Table 3

Average Results for 26-week Calibration Period

	National Launches – 12 Panel Tracking Data sets		Test Markets (HFW) – 19 BehaviorScan Data sets	
Model	Forecast MAPE	52w_APE	Forecast MAPE	52w_APE
Exp.-Gamma	9	9	10	12
Exp.-Gamma w /NT	9	9	9	12
Exp. w/NT	9	10	12	18
Mean	***9***	***10***	***10***	***14***

Table 4

Analogous Series of Cumulative Trial

	Proportion of Year-End Trial (Analogous Series)		52w_APE	
Week	National Launches	Test Markets	National Launches	Test Markets
4	0.05	0.32	84	33
8	0.15	0.47	57	24
12	0.25	0.58	26	22
***13**	***0.31***	***0.60***	***13***	***23***
16	0.35	0.64	18	22
20	0.45	0.70	11	18
24	0.56	0.74	10	14
***26**	***0.62***	***0.78***	***6***	***11***
28	0.64	0.80	7	9
32	0.72	0.84	9	7
36	0.78	0.88	7	6
40	0.85	0.91	4	3
44	0.89	0.94	4	2
**48	0.95	0.98	5	1
52	1.00	1.00	0	0

*US national launches interpolated for these periods due to missing data

**One New Zealand launch interpolated for this period due to missing data

Figure 1

Analogous Series for Three Data Sets

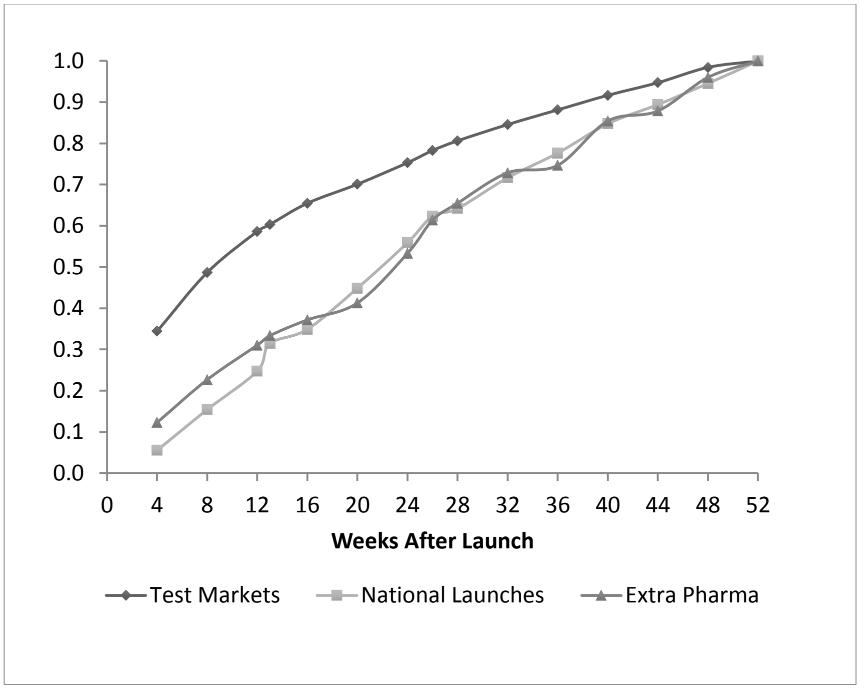

Figure 2

Normalized Trial for 34 Test Data Sets

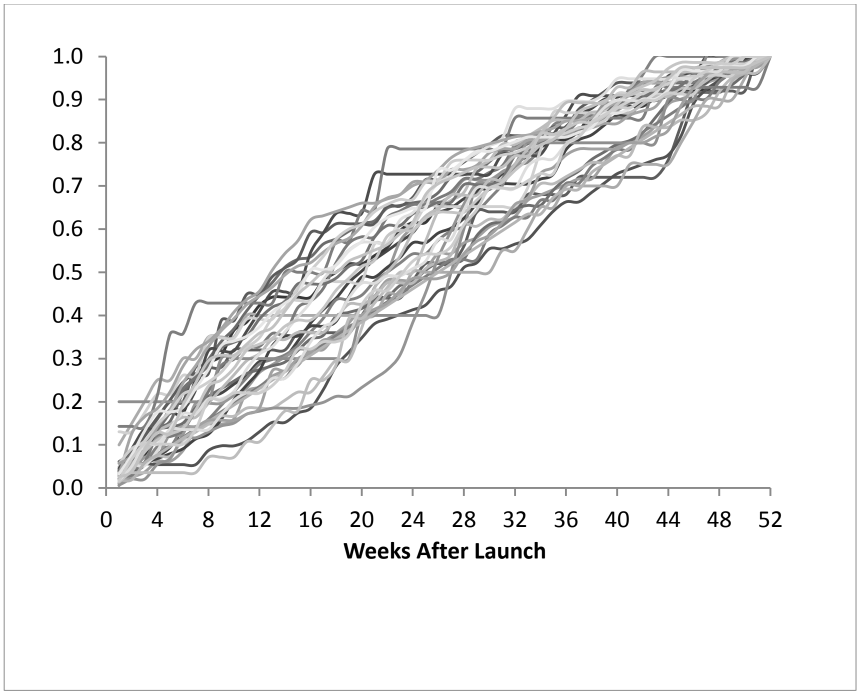

Table 5

Comparative Results for Additional 34 Test Data Sets

	13-week Calibration Period			26-week Calibration Period		
	Forecast MAPE	52w _APE	52w _MPE	Forecast MAPE	52w _APE	52w _MPE
12 National Launches - Exp.-Gamma	27	28		9	9	
19 BehaviorScan - Exp.-Gamma	17	23		10	12	
34 Test Data Sets – Straight Line	25	41	-33	13	25	-23
34 Test Data Sets – Exp.-Gamma	20	28	7	12	15	1
34 Test Data Sets - Analogous Series	*12*	*26*	*-6*	*7*	*13*	2