10-2008

Real-Time Evaluation of Email Campaign Performance

Andre Bonfrer
Singapore Management University, andrebonfrer@smu.edu.sg

Xavier Dreze
Wharton School, University of Pennsylvania

Real-Time Evaluation of Email Campaign Performance

October 2006

André Bonfrer*
Xavier Drèze

* Corresponding author. André Bonfrer is Assistant Professor of Marketing at the Singapore Management University, Lee Kong Chian School of Business, 50 Stamford Road #05-01, Singapore 178899, andrebonfrer@smu.edu.sg, Tel +65-6828-0398, Fax +65-6828-0777. Xavier Drèze is Assistant Professor of Marketing at the Wharton School of the University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104-6340, 215-898-1223, Fax 215-898-2534, xdreze@wharton.upenn.edu. This research was funded in part by the Wharton-SMU research center, Singapore Management University and in part by a WeBI - Mac Center Grant.

Real-Time Evaluation of Email Campaign Performance

We develop a testing methodology that can be used to predict the performance of email marketing campaigns in real time. We propose a split-hazard model that makes use of a time transformation (a concept we call virtual time) that allows for the estimation of straightforward parametric hazard functions to generate early predictions of an individual campaign's performance (as measured by open and click rates). We apply our method to 25 email campaigns and find that the method is able to produce in less than two hours estimates that are more accurate and more reliable than what the traditional method (doubling time) can produce after 14 hours.

Other benefits of our method are that we make testing independent of the time of day and we produce meaningful confidence intervals. Thus, our methodology can be used not only for testing purposes, but also for live monitoring. We show that a campaign selection rule based on our model rather than on the doubling method can improve overall response rates by 20%.

Keywords: Database marketing, email, pre-testing, advertising campaigns.

Introduction

Email can be a powerful vehicle for marketing communications. Many marketers favor this new medium because it provides them with a cheaper and faster way to reach their customers. Further, the online environment allows marketers to measure consumers' actions more accurately. This is a boon for marketing scientists in their desire to increase the effectiveness of marketing efforts and measure the ROI of marketing expenditures.

Although email response rates started out high (especially when compared with those reported for online and offline advertising), they declined over time and are now below 2.5% (DMA 2005). Finding ways to raise these response rates is critical for email marketers. A useful tool to achieve this is an effective email testing methodology. Identifying potential strengths and weaknesses of the content (the email creative) and the target population *before* the email is sent out at full scale can help marketers improve the response rates for their campaigns.

As a motivating example for the problem we are interested in, consider the case of a product manager at one of the major movie studios. With two or three new DVDs coming out every week, studios often rely on email marketing to generate interest for upcoming titles. This online promotion is particularly important for smaller titles (e.g., Transporter 2) that will not benefit from mass advertising and will not be heavily pushed by Amazon or Blockbuster. For such titles, the product manager would typically ask her creative agency to come up with a few concepts, and pick one of the designs to send out. She would also use a series of criteria (e.g., movie genre, gender) to select a subset of her database as a target for the email. Given the number of new titles to promote every week, she would work on a short production cycle and send emails without formally testing the quality of the design or the target sample (this is different from large releases such as X-Men III which are planned months in advance and

receive broad advertising and channel support). The success of our manager's emails might be much improved if she could test multiple creatives and target selection in a fast (she has short lead times) and inexpensive (the titles do not warrant large expenses) way. There are no models in the extant marketing science literature that can be directly applied to provide such a test.

The importance of testing elements of the marketing mix is not new to marketing scientists. For example, in new product development (and distribution) the ASSESSOR model (Silk and Urban 1978, Urban and Katz 1983) has been used for decades to forecast the success of new products based on laboratory based test marketing. Methods have also been developed to perform multiple parallel testing as predicated by the new product development literature (Dahan and Mendelson, 2001; Smith and Reinertsen, 1995). A novel approach used by Moe and Fader (2002) utilizes advance purchase orders made via the CDNOW website to generate early indicators of new product sales for music CDs. In advertising, the efficacy of an advertising campaign is assessed using a battery of tests designed to identify the best creative to use (e.g., the most persuasive or the most memorable) using selected members of the target audience. Field experiments with split-cable television technology have also been used to study the impact of advertising on brand sales (Lodish et al 1995, Blair 1988).

In direct marketing, modeling techniques have been developed to help marketers select the right customer group for a given content (Gönül and Shi 1998, Bult and Wansbeek 1995; Bitran and Mondschein 1996, Gönül, Kim, and Shi 2000, Gönül and Hofstede 2006). Bult and Wansbeek (1995) build a regression model to predict each customer's likelihood of responding to a direct marketing communication, and then select which customers should be contacted by explicitly maximizing the expected profits generated by each communication. Using a dynamic programming approach rather than using regression, Bitran and Mondschein (1996) take the

profit maximization a step further by incorporating inventory policies (inventory and out of stock costs) into the decision. Gönül and Shi (1998) extend this dynamic programming approach by allowing customers to optimize their own purchases behavior over multiple periods (i.e., both the firm and the customer are forward looking). In Gönül, Kim, and Shi (2000), the authors recognize that customers can order from old catalogs and that one can still garner new sales from old customers who were not sent a new catalog. Thus, they propose a hazard function model of purchase where customers are sent a catalog only if the expected profits with the additional mailing exceeds the profits without the mailing. Elsner, Krafft, and Huchzermeier (2004) use Dynamic Multilevel Modeling (DMLM) to optimize customer segmentation and communication frequency simultaneously.

In practice, few direct marketing campaigns are rolled-out untested. The traditional approach (Nash 2000) is to use the *doubling method* to predict the ultimate response rate to a direct marketing offer. In the doubling method, one uses past campaign to estimate the time needed for half of the responses to be received (the doubling time). Then, when performing a test, one waits the doubling time, and multiplies by two the number of responses received at that time to estimate the ultimate response rate. The waiting period depends on the medium used. For first class mailing, firms will wait a minimum of two weeks; for third class mailing, they will wait four weeks. Once the tests results are in the decision maker needs to make a go/no-go decision based on profitability. Morwitz and Schmittlein (1998) show that when making such projections, managers typically do not sufficiently regress test response rates to the mean.

Testing is also popular in Internet marketing applications. Online advertisers track banner ad performance in real time to identify the appeal (click-through) of various advertising creative. Click-stream models can be implemented to test the appeal of content by measuring the

click-through rates or website *stickiness* (Bucklin and Sismeiro 2003). Eye tracking technology may be used to identify where (and if) a customer is viewing the advertising message embedded on a webpage (Drèze and Hussherr 2003).

While some of these testing methodologies might be adapted to the context of email marketing, some unique features of email presents several new modeling challenges. First, many firms have implemented tracking technologies for email campaigns that can accurately measure (to the second) when a customer responds to the email. Given the goal of real-time testing, it is essential that we make full use of this continuous time data. The data tells us both whether and when a customer responds to an email, and we need our methodology to make use of this information pertaining to campaign level success.

Second, in contrast to a typical clickstream setting, email communications are initiated by the firm rather than the customer. This adds a layer of complexity in that, while the delivery of an email is often close to being instantaneous, there is an observable delay between the time the email is sent out and the time it is opened. The opening of the message will depend on how often customers check their email. Thus, although the marketer has direct control over when the email is sent out, there is little control over whether and when a customer responds to the email. This is different from traditional clickstream models where the user requests the content, and we can assume that the content is being processed immediately.

A third difference in email marketing involves the lead time for generating both the creative and the execution of the campaign. Even a large email campaign can be sent out at relatively low cost and delivered in a matter of hours, consequently campaigns are short lived and often run with short lead times and consequently with compressed deadlines such as weekly (e.g., American Airlines, Travelocity, The Tire Rack), bi-weekly (e.g., Longs Drugstore on the

West Coast), or even daily basis (e.g., Sun Microsystems). These short lead times place significant constraints on testing, simply because answers to a test are typically needed in just hours to be useful.

For these reasons, effective email marketing communication requires a testing methodology that is able to generate actionable results in as short a time as possible. The goal of such a testing procedure is to generate predictions of open incidence and click-through rates of any email campaign as quickly and accurately as possible. Our paper describes the development and test of an email pre-testing model. We begin by developing a split-hazard log-logistic model of open behavior. The split-hazard component models the incidence of open (versus not open); a Log-Logistic hazard rate is used to predict the distribution of open times. Click behavior is then modeled using both a censored split hazard model and a simpler Binomial model. To help us produce stable estimates even when data is sparse (a common occurrence when trying to test campaigns in a short amount of time) we use Bayesian shrinkage estimation. This allows us to take advantage of the information contained in past campaigns and weight this past information with response data observed in the focal campaign. Both sets of models are compared with the doubling method used by direct marketing practitioners.

In our application of the models to actual data, we find it necessary to account for intraday variations in customer responses (e.g., to account for fewer emails opened at night). Consequently, we develop a concept of virtual time that allows us to produce a model that fits the data well while keeping the specification simple. Virtual time involves adjusting the speed of time to adapt to the marketers' and customers' availability throughout the day. Using virtual time allows us to keep the model specification simple. This makes shrinkage straightforward and allows for an easy interpretation of the model parameters.

We apply the testing procedure to data obtained from a large entertainment company. The analysis shows that using our approach, we can reduce testing time from 14 hours to less than two without any decrease in testing reliability. It also highlights the pitfalls inherent with compressing testing time. Indeed, the more compressed the testing time, the more sensitive the quality of the results are to the specifications of the hazard function and to intra-day seasonality.

Our model provides a number of substantial practical benefits. (1) The fast evaluation of a campaign allows for early warnings about the probable success or failure of the campaign. This can lead to timely go/no-go decisions for either campaigns or creative. (2) Our model provides diagnostic information that can help improve the results of an under-performing campaign. A simple decision model could discard any campaign that does not perform above some threshold level of response. (3) Our testing procedure coupled with such a decision process can generate higher average (cross-campaign) response rates (20% higher in our simulation). (4) An important additional advantage of our testing procedure is that only a small sample is required for testing. The small sample size makes it easy to test the effectiveness of multiple advertising copies. Several sub samples can be generated, and a different creative sent to each sub sample. (5) Our process formalizes the use of the company's knowledge base in modeling future campaigns. Indeed, as the number of campaigns grows, the email marketer learns more about the distribution of response rates for campaigns sent. This leads to more accurate forecasts.

1. Research Setting and Data Description

We will calibrate and test our models using a database of twenty-five email campaigns sent as part of the online newsletter of an entertainment company. Most of the emails are in the form of promotions aimed at inducing customers to purchase a movie title online or offline, or to click on

links to access further content (different campaigns have different purposes). Each campaign has a subject line displaying the purpose of the promotion. The main body of the email is only visible after the recipient has opened the email. Within the body of the email, recipients are able to click on various links to activate the promotion or to direct them to a website. It is important to note that clicks can only occur if a recipient opens the email.

Summary statistics for the email campaigns are reported in Table 1. The campaigns vary in size from around 5,000 to 85,000 emails sent. Our database consists of 617,037 emails sent, of which 111,419 were opened, and 9,663 of those emails were clicked on at least once. The open rate is thus 18.1% and the click-through rate is 8.7%. The unconditional click rate (defined as the number of clicks divided by the number of emails sent) is about 1.6%.

There is a wide range in both open and click-through rates across campaigns. Clearly, the more successful campaigns are the ones that have both high open rates and high click-through rates. Using a median split on open and click-through rates, we find that 20% of our campaigns fall in this upper right quadrant (high open and click-through rates). Some campaigns (32% of our data) enjoy high open rates but have low click-through rates. This is probably an indication that these campaigns have broad appeal but are poorly executed. The firm might be able to move these campaigns to the upper right quadrant by using better creative. Another 28% of campaigns are in the opposite quadrant; they fail to attract enough attention to be opened, but generate high click-through given they were opened. In such cases, the execution is good, but the base appeal is low; the firm might improve the campaign through better targeting. The remaining campaigns (20%) have both low open and click-through rates; improving these campaigns would require improving both the targeting and the content. If this cannot be done, it may be best to drop the campaign altogether.

Given the number of campaigns in each quadrant, it is clear that it is difficult to predict the success of a campaign ex ante. The goal of our model is thus to predict out-of-sample open and click rates quickly and with small samples. Providing a forecast in a timely manner allows the firm to adjust the creative or targeting of the campaign when needed, thereby improving overall results.

Figures 1a and 1b present histograms of the time (in hours) for customers to open the email since it was sent, and the time (in minutes) it takes the customer to click on the email since it was opened. Given our objective of reducing the time allocated to testing, several features of our data are highly pertinent to model construction:

1. Opens usually occur within 24 hours of sending; clicks occur within a minute of being opened.
2. There is a relatively low level of email activity during the first few hours after an email campaign is sent (see the first few hours of Figure 1a), followed by a build up.
3. The histogram of the delay between send and open (Figure 1a) reveals a distinct multi-modal pattern underlying particularly during the first 24 hours after an email is sent. This pattern is also visible on individual campaign histograms.

The first feature requires that a rapid testing model of open and click rate work well with censored data. Indeed, by shortening the testing time, we reduce the amount of uncensored data available to us. The second feature suggests that we must be careful about our assumptions regarding the data generation process when building our model. This is particularly important in our case as we are trying to make predictions about the entire distribution of email activity based on only a few hours of activity.

The multimodal pattern found in the time until opening is troublesome as it does not appear to conform to any standard distribution and might be difficult to capture with a simple model. To understand what may be driving this multi-modal pattern, we plot the distribution of opens throughout the day (see Figure 2). This graph shows considerable variation in activity through the day. There are fewer emails opened late at night and early in the morning than during the day. We refer to this pattern as intraday seasonality. We show in the next section how this seasonality is the cause of the multimodal feature of Figure 1a.

Given these results, we will accommodate the following features in our modeling framework. To generate estimates within the first few hours after sending, we will have to work with censored data and only a small amount of data will be available for estimation. We also need to take into account intraday seasonality to allow parsimonious parametric approach to model the number of opens. In the next section, we develop a methodology that accommodates these issues.

2. Model Setup

We develop a rapid testing methodology for a specific application: the testing of online email campaigns. The rapid testing methodology requires a model that provides early feedback on whether a campaign is likely to be successful or not. In the spirit of traditional testing models, it is important that our methodology consumes as few resources as possible. Ideally, the model would also be parsimonious (i.e., have few parameters), and would estimate quickly such that a test can be implemented in real time and would allow for the monitoring of an email campaign as it is being sent. Indeed, an overly complex or over-parameterized model that takes hours to generate predictions would defeat the purpose of rapid testing.

We first describe in more detail how the model accommodates intra-day seasonality. Next, we develop a split hazard model of open and click probabilities that takes into account the possibility that some emails are never opened or clicked on (given open). We then derive the shrinkage estimators for the open and click models and state the likelihood function used in the estimation.

2.1. From physical time to virtual time

A traditional approach to handling seasonality, such as that displayed in Figures 1 and 2, is to introduce time-varying covariates in the model. There are two main problems with this approach. First, the covariates are often ad-hoc (e.g., hourly dummies). Second, they often make the other parameters less interpretable (e.g., a low open rate during peak hour could be larger than a high open rate during off-peak hours). To alleviate these concerns, we build on the approach developed by Radas and Shugan (1998). Radas and Shugan (hereafter RS) de-seasonalized a process by changing the speed at which time flows. They showed that by speeding up time during high seasons, and slowing down time during low seasons, one can create a new (virtual) time series that is devoid of seasonality. The benefits of this approach, assuming that one has the right seasonality pattern, is that one can use straightforward models in virtual time and easily interpret the meaning of the parameters of these models.

The effectiveness of the RS approach hinges on having a good handle on the seasonal pattern present in the data. In their application (the movie industry) they produce seasonal adjustments by combining past sales data with industry knowledge (e.g., presence of major holidays with high movie demand). A shortcoming of this approach is that some of the seasonality may be endogenous to the firms' decisions. For instance, if movie studios believe that Thanksgiving weekend is a 'big' weekend, they may always choose to release their best

movies during that weekend (Ainslie, Drèze, and Zufryden 2005). Thus, part of the seasonality observed during Thanksgiving will be due to the fact that more consumers have the time and desire to see movies on that weekend (consumer induced seasonality) and part of the seasonality will be due to the fact that firms release their bigger movies on that weekend (firm induced seasonality). If one uses past data as a base for seasonal adjustment without considering the decisions of the firm, one can potentially overcorrect and attribute all the seasonal effects to consumer demand while it is in fact also partly due to firm supply.

In our case, we also have a potential for both consumer- and firm-induced seasonality. For instance, the average consumer is much less likely to open emails at four in the morning than at four in the afternoon. Similarly, firms do not work 24 hours a day. If we look at when the firm sends its email (Figure 3), we observe little (but some) activity during the night, then a peak at eight in the morning, a peak at noon, an a lot of activity in the afternoon. It is likely that these peaks are responsible for some of the increase in activity we see in Figure 1 at similar times.

To separate consumer induced seasonality from firm induced seasonality, we benefit from two features of our modeling environment not present in RS. First, we have continuous time individual level data. While RS had to work with aggregate weekly measures, we know the exact time each email is sent and opened. Second, while a movie can open on the same day throughout the country, emails cannot all be sent at the same time. Emails are sent sequentially; for example, a million-email campaign can take 20 hours to send. Thus, we can simulate an environment that is devoid of firm based seasonality by re-sampling our data such that the number of emails sent at any point in time is constant through the say (i.e., Figure 3 for such a firm would be flat).

To resample the data, we proceed in three steps. First, for each minute of the day, we collect all emails that were sent during that minute. Second, we randomly select with replacement 100 emails from each minute of the day (144,000 draws). Third, we order the open times of these 144,000 emails from 0:00:00 to 23:59:59 and associate with each actual open time a virtual time equal to its rank divided by 144,000. The relationship between real and virtual time based on their cumulative density functions is shown in Figure 4. This represents the passing of time as seen by consumers independent of the actions of the firm.

We can use the relationship depicted in Figure 3 to compute the elapsed virtual time between any two events. For instance, if an email were sent at midnight and opened at two in the morning, we would compute the elapsed virtual time between send an open by taking the difference between the virtual equivalent of two a.m. (i.e., 00:29:44 virtual) and midnight (i.e., 00:00:00 virtual) to come up with 29 minutes and 44 seconds. Similarly, if the email had been sent at noon and opened at two p.m., then the elapsed virtual time would be 11:05:10 – 09:08:30 = 1 hour 56 minutes and 40 seconds.

Applying the virtual time transformation to the elapsed time between send and open for all emails in our dataset results in the histogram shown in Figure 4. Comparing this histogram to Figure 1, we can see the effect of using a virtual time transformation. The underlying seasonal pattern has all but disappeared. What was a multimodal distribution is now unimodal.

2.2. A split-hazard model of open and click time

The time it takes for customers to open an email from the time it is sent, or the time it takes to click on an email from the time the customer opens it are both modeled using a standard duration model (e.g., Moe and Fader 2002, Jain and Vilcassim 1991). Since both actions can be modeled using a similar specification, we discuss them inter-changeably. Starting with opens, we account

for the fact that in an accelerated test, a failure to open an email is indicative of one of two things. Either recipients are not interested in the email, or they have not had a chance to see it yet (i.e., the data is censored). Of course, the shorter the amount of time allocated to a test, the higher the likelihood that a non-response is indicative of censoring rather than lack of interest. To account for this bias, we model the open probability and the open time simultaneously in a right-censored split hazard model (similar to Kamakura, Kossar, and Wedel (2004) and Sinha and Chandrashekaran (1992)).

The probability that a customer will open or click an email varies from campaign to campaign, and is denoted with δ_e^k, where *e* is a subscript identifying different campaigns, and *k* is a superscript denoting an open (δ_e^o) or click (δ_e^c).

The likelihood function is constructed as follows. We start with a basic censored hazard rate model of the open time distribution:

$$L_e^k = \prod_{i=1}^{N_e} f\left(t_{ie}^k \mid \Theta_e^k\right)^{R_{ie}^k} S\left(T_{ie} \mid \Theta_e^k\right)^{1-R_{ie}^k}, \qquad (1)$$

where: e is a subscript that identifies a specific email campaign,

k is a superscript that identifies the model used ($k \in \{o = \text{open}, c = \text{click}\}$),

i is an index of recipients,

N_e is the number of recipients for email *e*,

R_{ie}^k is 1 if recipient *i* opened/clicked email *e* before the censoring point T_e,

T_{ie} is the censoring point of email *i* of campaign *e*,

t_{ie}^k is the elapsed time between send and open (open and click) in the event that the recipients opened (clicked) the email,

$f(t \mid \Theta)$ is the pdf for time *t*, given a set of parameters Θ,

$S(t \mid \Theta)$ is the corresponding survival function.

We adjust the hazard rate to account for the fact that some recipients will never open the email. If we call δ_e^o the likelihood that email *e* will be opened (δ_e^c for clicks), we have:

$$\begin{aligned} L_e^k\left(t_e^k, T_e, R_e^k \mid \Theta_e^k\right) &= \prod_{i=1}^{N_e}\left[\delta_e^k f(t_{ie}^k \mid \Theta_e^k)\right]^{R_{ie}^k}\left[\delta_e^k S(T_{ie} \mid \Theta_e^k) + (1-\delta_e^k)\right]^{(1-R_{ie}^k)} \\ &= \prod_{i=1}^{N_e}\left[\delta_e^k f(t_{ie}^k \mid \Theta_e^k)\right]^{R_{ie}^k}\left[1-\delta_e^k(1-S(T_{ie} \mid \Theta_e^k))\right]^{(1-R_{ie}^k)}. \end{aligned} \tag{2}$$

The estimation of δ_e^k and Θ_e^k for any parametric hazard function can be performed by maximizing this general likelihood function.

2.3. Shrinkage estimators

As in most practical applications, we benefit by having data available from past campaigns and we can use this information to improve the performance of our model. Specifically, we can use parameters from past campaigns to build a prior on the open and click hazard functions, as well as the split hazard component. This is especially useful at the beginning of a campaign when data is sparse.

The implementation of the shrinkage estimator depends on the specific hazard functions used in the model. We therefore postpone our discussion of the shrinkage estimators until the empirical section of the paper where we evaluate different possible hazard functions.

2.4. An Alternative Approach to Estimating Click Rates

Although theoretically sound, using a split-hazard model to estimate the parameters of the click times (conditional on an open) might be overly complex. Indeed, since most consumers click on an email within seconds of opening it, it is likely that few click observations are right-censored. In our sample, over 95 percent of all emails opened before the doubling point are also clicked on

before the doubling point. Therefore, we develop and test an alternative model for estimating click rates. We use a traditional binomial process with a Beta distributed prior in an empirical Bayes framework. Our hope is that a more parsimonious model will perform better at the beginning of a test, when few data points are available.

Formally, we use the mean of the posterior of the Beta distribution as the estimate for δ_e^c as follows. Let *c* be the number of clicks, *o* be the number of opens, and Beta(υ, ω) be our prior on the distribution of the click rate (built using prior campaigns). Then the posterior distribution of the conditional click likelihood is distributed Beta$(\upsilon + c, \omega + o - c)$. The estimate for the posterior mean is:

$$\hat{\delta}_e^c = \frac{\upsilon + c}{\upsilon + \omega + o} \tag{3}$$

As before, estimates of the prior distribution parameters (υ, ω) are generated from available past campaigns. The values for open (*o*) and click (*c*) are generated from the data at the time the test is conducted. We refer to this click model as the Empirical Bayes Binomial Distribution (EBB).

2.5. The Doubling Method

Before discussing an application of our model, we would like to draw a comparison with existing approaches to predicting the success rate of direct marketing campaigns. The most common model used by practitioners is the *doubling method* (Nash 2000). This method involves first examining the responses of past direct marketing campaigns and computing the amount of time it takes for 50% of the responses to be received (the *doubling time*). The analyst then uses the heuristic that for any future campaigns, the predicted total number of responses is equal to double the number of responses observed at the doubling time. In our case, the doubling time is 14 hours, ranging from 4 to 29 hours (see Table 1).

The doubling method is a powerful and simple heuristic. It makes three implicit assumptions. First, it assumes that not everybody will respond. Second, it assumes that it takes time for people to respond. Third, it assumes that the timing of the responses is independent from the rate of response and approximately constant across campaigns. As a non-parametric method, it does not make any assumption about the underlying response process, nor does it provide any ways to test whether the current campaign conforms to the data collected from previous campaigns or runs faster or slower than expected. Hence, it does not provide any ways to evaluate whether a current test should be run for a longer period or could be finished early; an important piece of information our technique provides.

In essence, the doubling method aggregates time into two bins; each containing half of the responses. This aggregation loses vital timing information that could be used to better model the response process.

3. Application of the model to email campaign pre-testing

We now apply our models to the data from the email campaigns described earlier and evaluate the relative predictive validity of various models. Our application involves two main phases, a calibration and model specification phase and an estimation and validation stage. We compare the predictions of our models with those of the "doubling" heuristic often used in direct marketing applications.

Calibration and Model Specification: In the calibration phase, we are interested in learning which of the parametric functional forms we should use to drive our predictive models. To implement this we compare a set of commonly used hazard rate distributions including Weibull/Exponential, Log-Logistic and Log-Normal. The functional form that fits the data best

is used in the simulation and calibration phase. The specification chosen for the hazard function also drives the specification required for shrinkage estimation.

Simulation and validation: the main purpose of the simulation and validation stages is to validate the models proposed in the paper by studying the accuracy of the predictions they make out-of-sample. We also want to find out which of the models has the best predictive performance and whether we can generate estimates that are useful for decision making within a short amount of time (say hours) such that testing is feasible for real-time campaign planning.

Given the best specification for the hazard rate found in the calibration phase, several models are compared for both the open and the click processes. We compare the models based on real time (no time transformation) versus virtual time, and based on shrinkage versus no-shrinkage estimation. In summary, we fit and validate each of the following models for both the open and click processes in the simulation and validation phases:

1) No-shrinkage estimation with real time
2) No-shrinkage estimation with virtual time
3) Shrinkage estimation with real time
4) Shrinkage estimation with virtual time

The EBB model is tested only for the clicks and represents our fifth specification tested:

5) Empirical Bayes Binomial model for clicks only

There is no need to test the virtual time transformation for the EBB model because this specification uses only a count of the number of clicks relative to the number of opens – it is therefore not based on the amount of time it takes for customers to click on the email given it was opened.

3.1. Calibration and Estimation

Shape of the Hazard Function

Researchers in marketing have employed several specifications for the hazard function when doing survival analysis see Jain and Vilcassim 1991, Sawhney and Eliashberg 1996, Chintagunta and Haldar 1998, Drèze and Zufryden 1998). Following their work, we considered the following four specifications: Exponential, Weibull, Log-Normal, and Log-Logistic. This set of distribution encompasses a wide range of consumer behavior. Our final choice of hazard function is based on how well it agrees with the data (goodness of fit).

We estimated a campaign level hazard rate model for each distribution using the complete set of opens and clicks available for each campaign (i.e., this is a straight hazard model that is neither split nor censored). We report the fit statistics for all four specifications in Table 2a (open model) and 2b (click model). The analysis suggests that the Log-Logistic distribution fits the data best overall for both open and click. The Log-Normal is a close second, but has the drawback of not having a closed form expression for its survivor function. It is important to note that the Exponential distribution performs relatively poorly, emphasizing the need for a non-constant hazard rate that allows for a delay between reception and open of an email, or between open and click (i.e., allows for enough time for consumers to process the message). The relatively poor fit of the Weibull distribution (which allows for a ramping up period) further shows that one also needs to accommodate for a decrease in the hazard rate after enough time has passed. Making the right assumptions regarding the change in hazard rate over time is thus crucial. This is especially true since much of the data available during the test will come from the first few hours of the test, representing the increasing part of the Log-Logistic hazard function.

Estimating this based on a Weibull or Exponential hazard function would clearly misspecify the model.

The probability density function and the survivor function for the Log-Logistic are (see Kalbfleisch and Prentice (1985) for details about the Log-Logistic and other distributions mentioned in this paper):

$$f(t \mid \alpha, \lambda) = \frac{\lambda\alpha(\lambda t)^{\alpha-1}}{\left(1+(\lambda t)^{\alpha}\right)^{2}}, \qquad S(t \mid \alpha, \lambda) = \frac{1}{1+(\lambda t)^{\alpha}} \tag{4}$$

where $\lambda > 0$ is a location parameter and $\alpha > 0$ is a shape parameter. Consistent with previous notation, we refer to the shape and location parameters for any given campaign (*e*) and email response action ($k \in \{o, c\}$) as α_e^k, and λ_e^k, respectively. Depending on the value of α, the Log-Logistic hazard is either monotonically decreasing ($\alpha \leq 1$) or inverted U-shape ($\alpha > 1$) with a turning point at $t = \frac{(\alpha-1)^{1/\alpha}}{\lambda}$.

Shrinkage Methodology

Since we use a Log-Logistic hazard function, our split-hazard models have three parameters $(\alpha, \lambda, \delta)$. We build informative priors for $(\alpha, \lambda, \delta)$ using the estimates obtained from other campaigns. Based on an inspection of the empirical distribution of these parameters, we specify our prior for δ as a Beta distribution and α and λ are as a bivariate Log-Normal distribution:

$$\begin{aligned} \delta &\sim \text{Beta}(a_\delta, b_\delta) \\ (a, \lambda) &\sim \text{Log-Normal}(\mu, \Sigma) \end{aligned} \tag{5}$$

The parameters $(a_\delta, b_\delta, \mu, \Sigma)$ for a given campaigns are estimated using the method of moments from the parameters $(\alpha, \lambda, \delta)$ obtained from all other campaigns. To compute the correlation between parameters α and λ of the Log-Normal distribution we use the correction factor described in Johnson and Kotz (1972, page 20) to adjust for possible small sample bias. The correlation for the Log-Normal is:

$$\rho_{\alpha,\lambda}^{LN} = \frac{\exp(\rho_{\alpha,\lambda}^{N}\sigma_\alpha\sigma_\lambda)-1}{\sqrt{\left\{\exp\left(\sigma_\alpha^2\right)-1\right\}\left\{\exp\left(\sigma_\lambda^2\right)-1\right\}}} \tag{6}$$

where $\rho_{\alpha,\lambda}^{N}$ is the correlation coefficient for the Normal distribution of the two parameters. Note that this correlation coefficient is independent of the means of the parameters, depending only on the standard deviations of the parameters (respectively, $\sigma_\alpha, \sigma_\lambda$).

3.2. Simulation and validation

Split-Hazard Model

With the hazard function properly defined and the shrinkage methodology in place, we are now ready to fit and validate our models. The final likelihood functions are as follows:

1) Likelihood function for the Non-Shrinkage Model:

$$L_e^k\left(t_e, T_e, R_e^k \mid \Theta_e^k\right) = \prod_{i=1}^{N_e}\left[\delta_e^k \frac{\lambda\alpha(\lambda t_{ie})^{\alpha-1}}{\left(1+(\lambda t_{ie})^\alpha\right)^2}\right]^{R_{ie}^k}\left[1-\delta_e^k\left(1-\frac{1}{1+(\lambda T_{ie})^\alpha}\right)\right]^{(1-R_{ie}^k)} \tag{7}$$

2) Likelihood function for the Shrinkage Model

$$L_e^k\left(t_e, T_e, R_e^k \mid \Theta_e^k\right) = \mathrm{LN}\left(\mu, \Sigma\right)\mathrm{Beta}\left(a_\delta, b_\delta\right)\times \prod_{i=1}^{N_e}\left[\delta_e^k \frac{\lambda\alpha(\lambda t_{ie})^{\alpha-1}}{\left(1+(\lambda t_{ie})^\alpha\right)^2}\right]^{R_{ie}^k}\left[1-\delta_e^k\left(1-\frac{1}{1+(\lambda T_{ie})^\alpha}\right)\right]^{(1-R_{ie}^k)} \tag{8}$$

where Σ is the variance-covariance matrix for the Log-Normal prior distribution. The EBB estimator is computed directly from (3) without resorting to its likelihood function.

In each simulation, we adopt the perspective of a marketer who wishes to pretest his campaigns before committing to the final send. To this end, we look at each campaign assuming that the remaining (E - *1*) campaigns have already been completed. Thus, prior to the test, we know nothing about a focal campaign except the information contained in the priors, and the number of emails that need to be sent out. Based on company policy, we set our sample test size at 2,000 emails (we varied the test size between 1,000 and 2,000 in increments of 200 emails but did not find any substantive difference in results). We also set different censor points, in 30 minute increments, ranging from 30 minutes to 8 hours. At each censor point, any email that had been opened prior to the censor point was used in the non-censored component of the log-likelihood. All observations beyond the censor point (regardless of whether they were ultimately opened or not) were coded as censored. Based on this set of censored and uncensored observations, we then estimated the parameters of the split-hazard censored Log-Logistic rate model using priors constructed based on all other campaigns.

As a practical matter, the distributions of open and click experience long tails, such that some responses continue to come in long after a campaign has run its course. Our data reveals that 99% of all email responses are observed within 3 weeks of sending out the email communication. Typically, the company conducts post-campaign debriefing 2-3 weeks after the emails are sent out. Thus, we set a cut-off date of 3 weeks (504 hours) and use the numbers of opens and clicks observed at that time as the true value we are trying to predict. Our forecast of the number of opens and clicks at 3 weeks is constructed using the parameter estimates for each

of the censored samples. The cumulative distribution of opens and clicks at 504 hours is calculated using:

$\hat{\delta}_e^o \times F\left(504 \text{ hours} \mid \hat{\alpha}_e^o, \hat{\lambda}_e^o\right) \times \left(\text{Number Sent}_e\right)$ for the number of opens at 3 weeks, and

$\hat{\delta}_e^c \times F\left(504 \text{ hours} \mid \hat{\alpha}_e^c, \hat{\lambda}_e^c\right) \times \left(\text{Estimated opens}_e\right)$ for the estimated number of clicks.

Results for the "open" models

The full results for each campaign and for each censoring point consists of a set of parameters $(\delta_e^o, \alpha_e^o, \lambda_e^o)$ for opens and $(\delta_e^c, \alpha_e^c, \lambda_e^c)$ for clicks. Since for each model we have 16 time points (between 30 minutes and 8 hours) and 25 campaigns, this means our analysis generates a total of 6x16x30 = 2,880 estimates. Given this large number of estimates it is difficult to present all the results in one table. It is also not that meaningful to present any sufficient statistic of these estimates since they are censored at a different point in time.

We therefore summarize our results by looking only at the estimated open and click rate response rates for each campaign (δ_e^o, δ_e^c). Our summary includes a comparison of the predictions based on each set of estimates and at any given time point with their true (post-campaign) values. The summary also includes the corresponding average deviations of the predictions from their true values.

Figure 6 graphically presents this summary of the Mean Absolute Deviation (MAD) for the predicted number of opens across campaigns and in half hour increments. The vertical axis represents the MAD, across campaigns, or the average absolute difference between the predicted number of opens from the model and the actual number of opens observed in the data. The horizontal axis represents the half hour increments, starting at 30 minutes. The prediction from the 14 hour doubling point is also plotted on the graph, the MAD is indicated by the height of the

vertical line plotted at six hours. It is plotted at six hours because that is the time at which the best performing model yields an improvement over the doubling time model.

We observe a general downward trend for all models as the censoring point moves to the right and a larger proportion of the total sample is available for calibration versus prediction. The results show a clear dominance in predictive performance of the shrinkage models over the non-shrinkage models. The MAD values for the shrinkage models are about a quarter of the MAD values for the non-shrinkage models. Based on prediction error, the virtual time/shrinkage model tends to outperform all other models. Further, it achieves the same level of prediction error as the doubling method in as little as six hours compared to the 14 hours of the doubling method.

To get a better sense of the benefits and drawbacks of each model, we plot the predicted number of opens for an illustrative campaign for each model (Figure 7). The solid lines in the middle of each graph in Figure 7 represent the estimated number of opens at 504 hours, based on the censor point listed on the horizontal axis. The small dotted lines tracking above and below each of the solid lines represent the confidence intervals. The constant dashed line in the middle is the true value for the number of opens at 504 hours. We find that, for this campaign, the non-shrinkage models tend to perform quite badly compared with the shrinkage models. We see also that they have some difficulty converging especially within only a few hours of sending the emails.

A comparison across campaigns for each of these models reveals several benefits of the shrinkage over the non-shrinkage models. In short we find that non-shrinkage models exhibit three types of problems. First, they often fail to converge during the first couple of hours of

testing. Second, when they do converge, they often produce confidence intervals that are too tight. Third, they traditionally underestimate response rates during the first day of testing.

Shrinkage alleviates these problems, both in real time and in virtual time. Models converge quickly and reliably even with few data points. The confidence intervals produced are realistic. The benefits of the virtual time model over the real time model is that virtual time produces tighter and more stable estimates and produces the estimate faster. Thus the virtual time model is well suited to test the performance of a campaign.

Results for the "click" models

When predicting click activity in our simulations, we proceed in two steps. First, we fit our models of click-through rate (using equations (7), (8) or (3) as appropriate). We produce estimates for $\hat{\delta}_e^c$ (five estimates total, four for the log-logistic hazard rate models plus one for the EBB model). Second, we predict the ultimate numbers of click for each model by multiplying $\hat{\delta}_e^c$ with the number of opens ($\hat{\delta}_e^o$) forecast obtained from the virtual time/shrinkage model of opens at the same point in time. The virtual time/shrinkage model for open was used as a basis for this prediction because this is the one that performed best in the open model test and therefore represents the most realistic comparison.

We compare the estimated clicks with the actual clicks and compute the MAD of each model. Figure 8 reports MAD in 30-minute increments across all 25 campaigns for the four basic models and for the EBB model. The picture depicted here is similar to the case of the open prediction. The major difference is that the EBB model reveals itself to be the best performing model. Indeed, by combining the EBB model with the virtual time open model, we can achieve a performance similar to the doubling method in only three and a half hours–a 75% reduction in testing time.

3.3. A Campaign Selection Decision Rule

The preceding two sections show that our models can produce predictions of the open and click rates faster and more accurately than the doubling method. While the results do show a clear improvement in speed and accuracy, a natural question that arises is: are the benefits of the new models substantial from a managerial standpoint? To enumerate these benefits in the context of our application, we take the perspective of a campaign manager who uses a simple heuristic to eliminate any under-performing campaigns. Let us say that the manager only wants to run campaigns that will produce an unconditional click-through rate (CTR) of 2% or more. One possible reason for not wanting to send a low yield campaign is that it costs more to send than its expected returns. Such underperforming campaigns also represent an opportunity cost in that they tie up resources that could be used to send more profitable campaigns. Furthermore, sending undesirable material could lead to higher customer attrition.

We consider three different decision rules and compare the aggregate performance of the accepted campaigns under each decision rule. First, we use the doubling method, where we wait for 14 hours and select any campaign with a predicted CTR of 2% or more. Second, we use our best performing model and select any campaign with a predicted CTR of 2% or more after three and a half hours of virtual time regardless of the confidence interval around the estimate. Third, we use our best performing model and run the test until the predicted CTR is significantly different than 2% at $p = .05$. Using this third rule, the time needed to test a campaign will vary depending on the observed data.

The results of this test are shown in Table 3. As expected, our model performs better than the doubling method. The average campaign response rate can be increased by 9% simply by

applying our model and waiting for 3.5 hours of virtual time (3.63% CTR vs. 3.34%). One can gain a further 10% (4.00% vs. 3.63%, for a total improvement of 20%) by running the test to statistical significance rather than using a fixed time stopping rule. One should note that when using a statistical test rather than a rule of thumb to decide when to stop the test, the testing time varies widely across campaigns. Nine of the 25 campaigns reach statistical significance in as little as an hour. In contrast, one campaign takes the better part of three days (63 hours) before reaching statistical significance (it reaches 90% significance level after 90 minutes).

3.4. When Is the Best Time To Test?

Our comparison of the predictive ability for the split hazard rate model demonstrates that, on average, we can learn as much in three and a half hours as we can learn from the doubling method in 14 hours. However, it is important to remember that these three and a half hours are measured in virtual time. In real time, the test will require more or less time depending on when it is performed. Figure 9 shows how long three and half virtual hours correspond to in real time, depending on the time of day when the test commences. There appears to be a "sweet spot" in the afternoon, between 1pm and 7pm where a three and half virtual hour test can be carried out in much less than three actual hours (the shortest it could take would be 1:51 hours if it starts at 5:31 p.m.). Starting after 7pm will impose delays as the test is unlikely to be finished before people go to bed; if the test is started at 10:33 p.m. it will actually take 7 and a half hours to complete.

4. Discussion and Conclusion

The value of information increases with its timeliness. Knowing quickly whether a campaign is going to be successful provides the opportunity to correct potential problems before it is too late or even stop the campaign before it is completed. It is therefore imperative to develop methods

that improve both the accuracy and the speed with which campaign testing can be done. In this article, we study a modeling procedure that can be implemented for the fast evaluation of email campaign performance.

The performance of an email campaign is defined by its open and click rates. The methodology we propose predicts these rates quickly using a small sample pretest. Reducing the sample size and testing period to a minimum produces multiple modeling challenges. Indeed, we propose to send 2,000 emails, and wait less than two hours to produce estimates of how the campaign will perform after three weeks. In two hours, we typically observe fewer than a hundred opens and fewer than ten clicks. The key to successful prediction of the ultimate results of an email campaign based on so few data points lies in using the information to its fullest potential.

There are three elements that make our methodology successful: (1) using the appropriate model specification, (2) transforming time to handle intra-day seasonality, and (3) using informative prior information. Each of these three elements provides its own unique contribution to the overall fit of the model.

The appropriate hazard function is critical because our compressed-time tests produce observations that are heavily right censored. Thus, we are often fitting a whole distribution based only on its first quartile (or even less). A misspecification of the hazard function could cause severe errors in prediction. In other words, the value of the responses of the first few people to respond to the email campaign is an important indicator of the success of the overall campaign. We find that the traditional exponential hazard function used in many models of online behavior is a poor fit for our process. Our results provide strong evidence to suggest that email response rates (both open and click-through) are driven by a non constant hazard rate. Rather we see the

hazard rate rises in the early phase of an email campaign and then decreases as time progresses. We find that the best fitting parametric model for open times is the Log-Logistic.

For modeling click-through rates, we compare the same Log-Logistic hazard rate model with a Binomial process. We find that the straight binomial process is a good descriptor of the phenomenon given that the quick consumer response after an email is opened limits censoring to the point where it is not a factor. Thus, we find that the click-through rate (the total number of clicks for a campaign, unconditional on open) is best predicted using a combination of the Beta binomial model for the click rate, and the Log-Logistic split hazard model for the open rate.

We apply our split-hazard model to a virtual time environment. The virtual time transformation removes intra-day seasonality and makes our testing procedure invariant to the time of day at which it is performed. This is a key factor in the robustness of our model in that it allows us to bypass the need to handle seasonality directly in the model and allows for a straightforward specification with few parameters. By limiting the numbers of parameters we must estimate to three for the open model and one for the click model, we make the best use of our limited data (we have a high ratio of data points to parameters, or high degrees of freedom) and we produce parameters that are directly interpretable (the click and open rates or estimate directly without the need for transformation).

Another benefit of our time transformation is that by making each campaign independent of the time of day, we can compare results across campaign, and easily build informative priors for each of the parameters we need to estimate. This yields a procedure that produces meaningful estimates and confidence intervals with a minimum amount of data. It also allows a firm to conduct tests serially. That is, if they chose to modify a campaign's creative of target population as the result of a test, they can retest the campaign and compare the new results to the first ones.

Overall, by putting these three elements together, we are capable of running in 1:51 a test that produces similar results to a traditional test in 14 hours (an 85% decrease in testing time). In addition, our methodology produces meaningful confidence intervals. The implication of this is that the firm can monitor the accuracy of its test and decide to run the test for a longer or shorter period of time depending on how well it performs. This finding is particularly important when one considers that a million-email campaign can take up to 20 hours to send. Using our model one could monitor the campaign as it is being sent, and implement a "stopping rule." The rule would allow a manager to make a decision to terminate a campaign that is underperforming, or even to change the creative for the remaining recipients. Thus, our methodology can be used not only for testing purposes, but also for live monitoring. If done right, this could significantly improve average response rates by limiting the detrimental impact of poor performing campaigns.

Our model represents a first step towards better online marketing testing. As such, there is still much that can be accomplished. For instance, due to the lack of individual level information in our dataset, we did not include covariates in our models. It is likely that adding such information, were it available, should improve fit and predictive power. Further, if our dataset contained many data points per recipients (we have an average of 2 emails sent per name) one could model unobserved heterogeneity. Given current computational power such a model might, however, lead to estimation times that are too long to be beneficial in practice.

Another issue is the link between click and purchase behavior. The assumption is that click behavior is a measure of interest and is highly correlated to purchase. As campaign managers are evaluated based on clicks, we feel our analysis is appropriate. However, in future applications and with better data, it should be possible to link click and purchase behavior, and thus optimize purchases rather then clicks.

A final area for further research involves expanding our model to allow for a biviarate process between open and click responses. In the current model, we assumed independence between open and click probabilities. It may further improve predictions to allow for some dependence between these two processes and can be used to test the idea that recipients who open an email quickly may also be more likely to click on links an email.

References

Ainslie, Andrew, Xavier Drèze, and Fred Zufryden (2005), "Modeling Movie Life Cycles and Market Share," *Marketing Science*, 24 (3), 508-17.

Bitran, Gabriel D. and Susana V. Mondschein (1996), "Mailing Decisions in the Catalog Sales Industry," *Management Science*, 42 (9), 1364-81.

Blair, Margaret Henderson (1988) "An Empirical Investigation of Advertising Wearin and Wearout," *Journal of Advertising Research*, December 1987/January 1988, 45-50.

Bucklin, Randolph E. and Catarina Sismeiro (2003), "A Model of Web Site Browsing Behavior Estimated on Clickstream Data," Journal of Marketing Research, XL (August), 249-67.

Bult, Jan R. and Tom Wansbeek (1995), "Optimal Selection for Direct Mail," *Marketing Science*, 14 (4), 378-94.

Chintagunta, Pradeep K. and Sudeep Haldar (1998), "Investigating Purchase Timing Behavior in Two Related Product Categories," *Journal of Marketing Research*, XXXV (February), 43-53.

Dahan, Ely and Haim Mendelson (2001), "An Extreme-Value Model of Concept Testing," *Management Science*, 47 (1), 102-16.

Direct Marketing Association (2005), "The DMA 2005 Response Rate Report."

Drèze, Xavier and Francois-Xavier Hussherr (2003), "Internet Advertising: Is Anybody Watching?," *Journal of Interactive Marketing*, 17 (4), 8-23.

Drèze, Xavier and Fred Zufryden (1998), "A Web-Based Methodology for Product Design Evaluation and Optimization," *Journal of the Operation Research Society*, 49, 1034-43.

Elsner, Ralf, Manfred Krafft, and Arnd Huchzermeier (2004), "Optimizing Rhenania's Mail-Order Business Through Dynamic Multilevel Modeling (DMLM) in a Multicatalog-Brand Environment," *Marketing Science*, 23 (2), 192-206.

Gönül, Füsun and Frenkel Ter Hofstede (2006), "How to Computer Optimal Catalog Mailing Decisions," *Marketing Science*, 25 (1), 65-74.

Gönül, Füsun, Byung-Do Kim, and Meng Ze Shi (2000), "Mailing Smarter to Catalog Customers," *Journal of Interactive Marketing*, 14 (2), 2-16.

Gönül, Füsun and Meng Ze Shi (1998), "Optimal Mailing of Catalogs: A new Methodology Using Estimable Structural Dynamic Programming Models," *Management Science*, 44 (9), 1249-62.

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

Johnson, Norman L. and Samuel Kotz (1972), *Distributions in Statistics: Continuous Multivariate Distributions*. New York, Wiley.

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

Kalbfleisch, John D. and Ross L. Prentice (1985), *The Statistical Analysis of Failure Time Data*, John Wiley & Son, Inc., New York, Ney-York.

Kamakura, Wagner A, Bruce S Kossar, and Michel Wedel (2004), "Identifying Innovators for the Cross-Selling of New Products," *Management Science*, Vol 50(8), 1120-1133.

Moe, Wendy W. and Peter S. Fader (2002), "Using Advance Purchase Orders to Forecast New Product Sales," *Marketing Science*, 21 (Summer), 347-364.

Morwitz, Vicki G. and David C. Schmittlein (1998), "Testing New Direct Marketing Offerings: The Interplay of Management Judgment and Statistical Models," *Management Science*, 44 (5), 610-28.

Nash, Edward (2000), *Direct Marketing*, McFraw-Hill Eds, New-York, New-York.

Radas, Sonja and Steven M. Shugan (1998), "Seasonal Marketing and Timing Introductions," *Journal of Marketing Research*, 35 (3), 296-315.

Sawhney, Mohanbir S. and Jehoshua Eliashberg (1996), "A Parsimonious Model for Forecasting Box-Office Revenues of Motion Pictures," *Marketing Science*, 15 (2), 113-131.

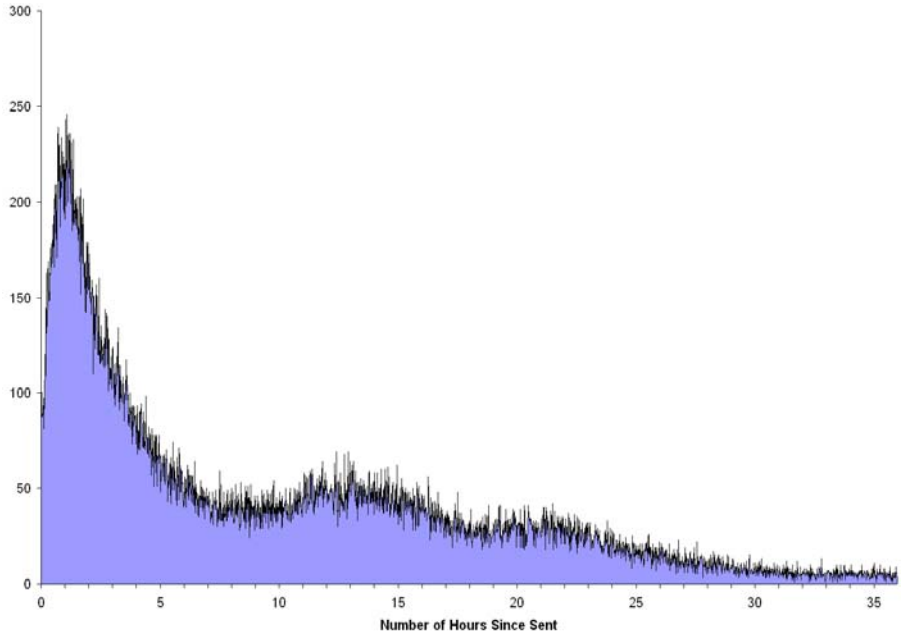
Silk, Alvin J. and Glen L. Urban (1978), "Pre-Test-Market Evaluation of New Packaged Goods: A Model and Measurement Methodology," *Journal of Marketing Research*, XV (May), 171-91.

Sinha, Rajiv V and Murali Chandrashekaran (1992) "A Split Hazard Model for Analyzing the Diffusion of Innovations," *Journal of Marketing Research*, Vol XXIX (February), 116-127.

Smith, Preston G. and Donald G. Reinertsen (1995), *Developing Products in Half the Time*, John Wiley & Sons, Inc. New York: New York.

Urban, Glen L. and G. M. Katz (1983), "Pre-Test-Market Models: Validation and Managerial Implications," *Journal of Marketing Research*, XX (August), 221-34.

Figure 1a and 1b. Histograms of elapsed time between sent and open (1a) and between open and click (1b) events, across all campaigns.



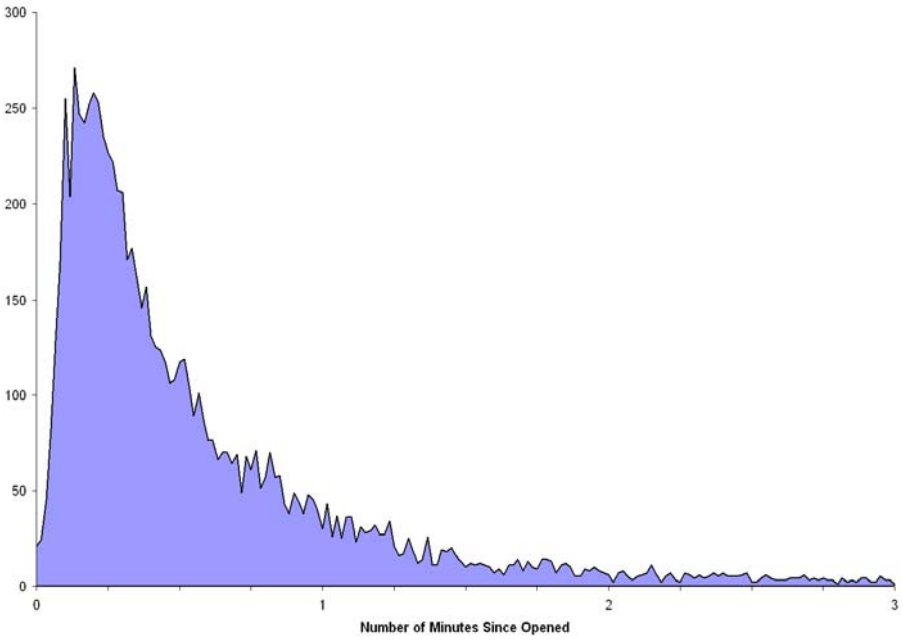


Figure 2: Distribution of open time through the day (Pacific Standard Time).

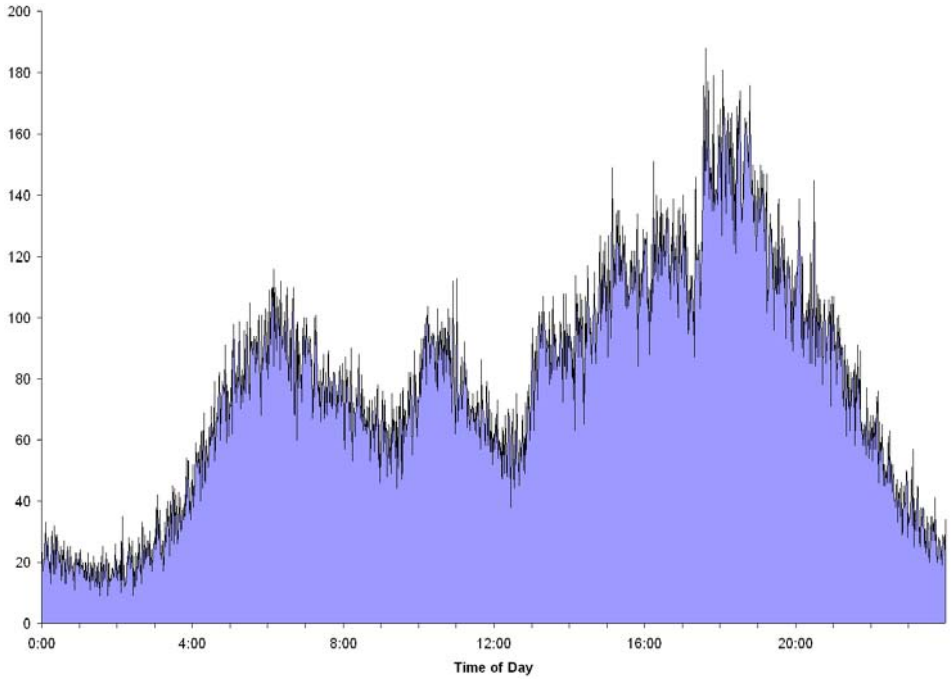


Figure 3: Distribution of Email Sent through the day

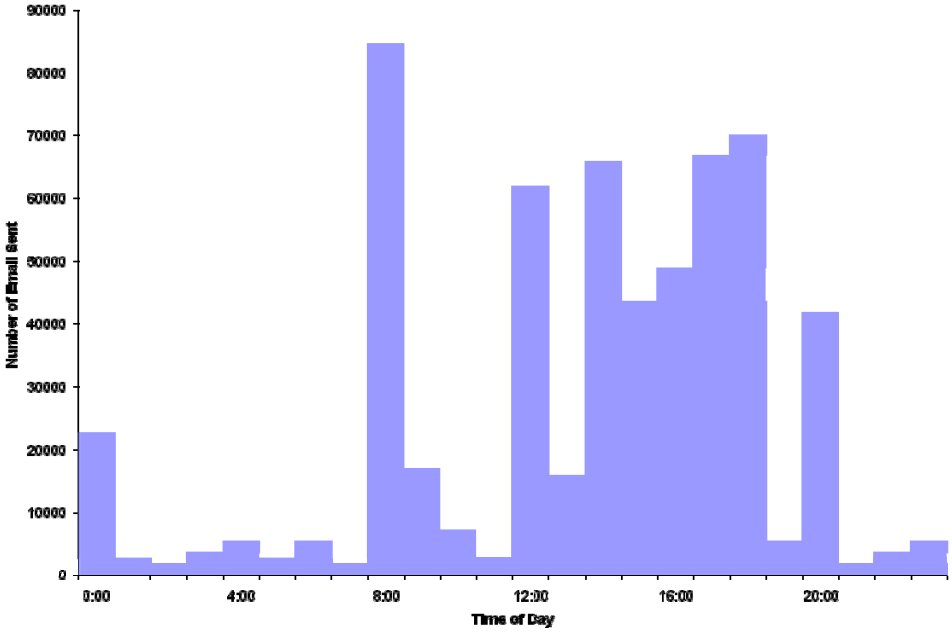


Figure 4: Within day cumulative density functions of real and virtual time. Real time is distributed uniformly throughout the day and therefore appears as the 45 degree line.

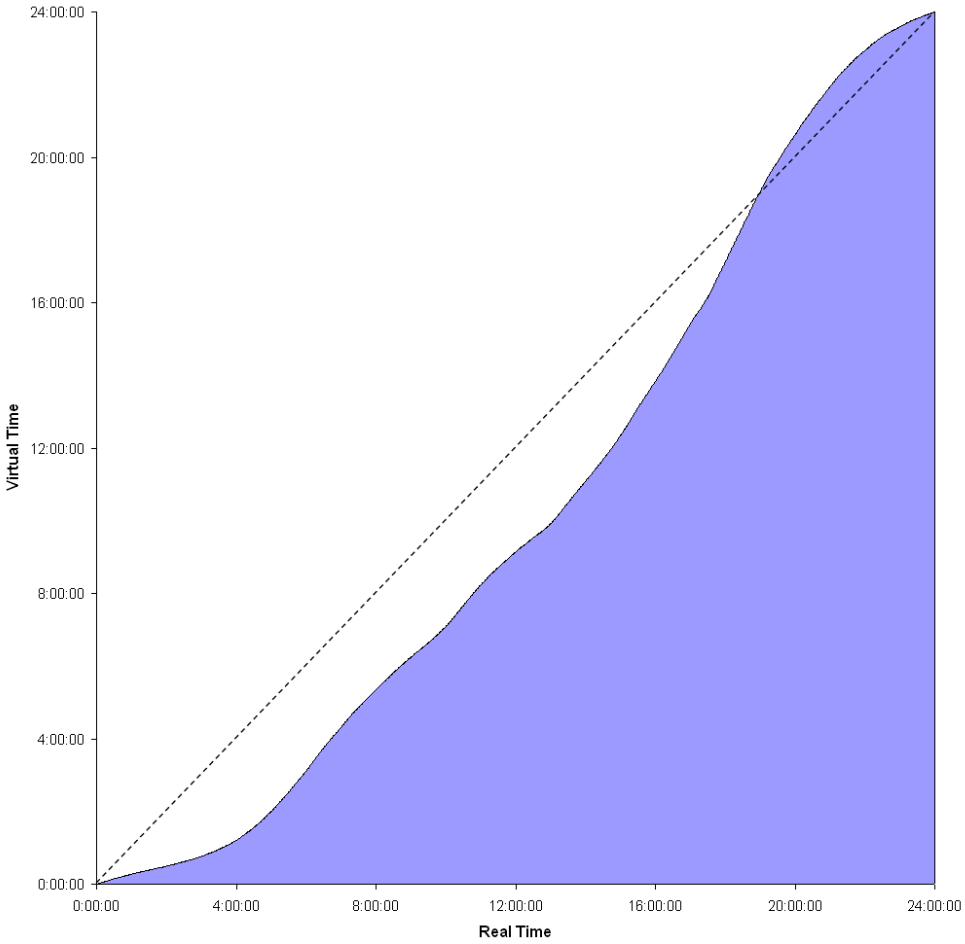


Figure 5. Histogram for virtual elapsed time between sent and open across all campaigns.

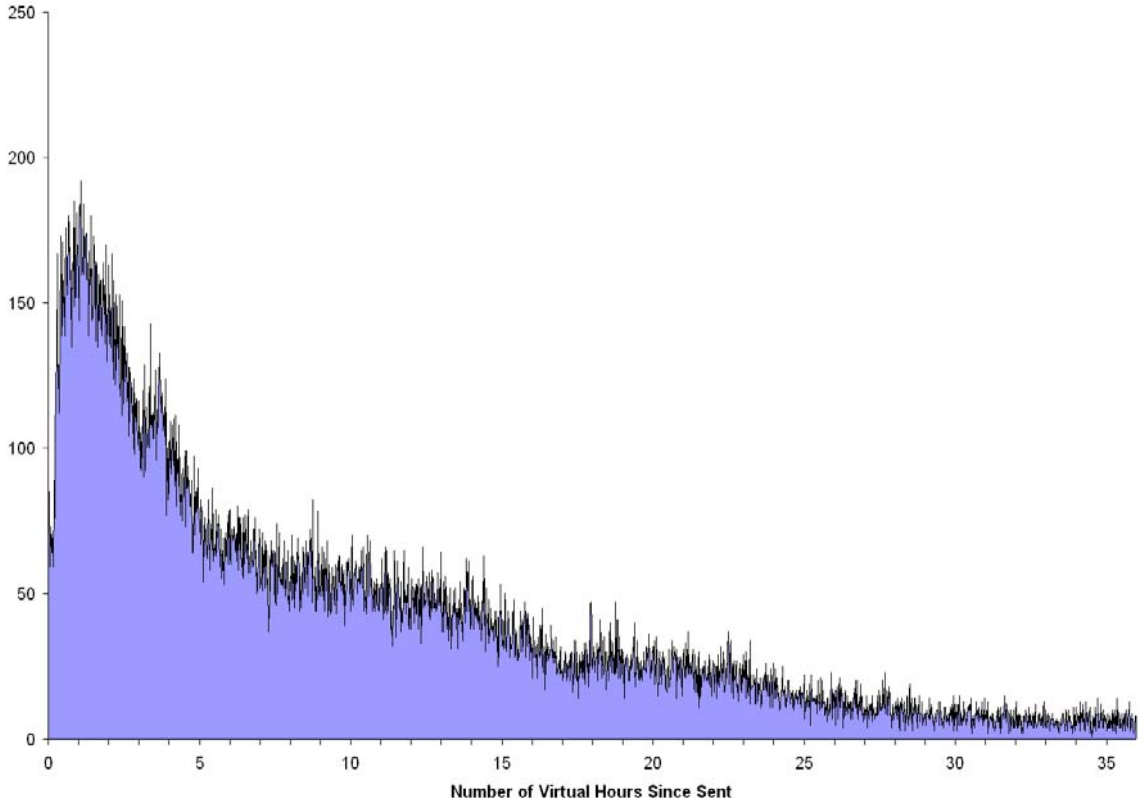


Figure 6: Mean Absolute Deviations (MAD) across models of Open response. The height of the vertical dotted line represents the (14 hour) MAD for the doubling method.

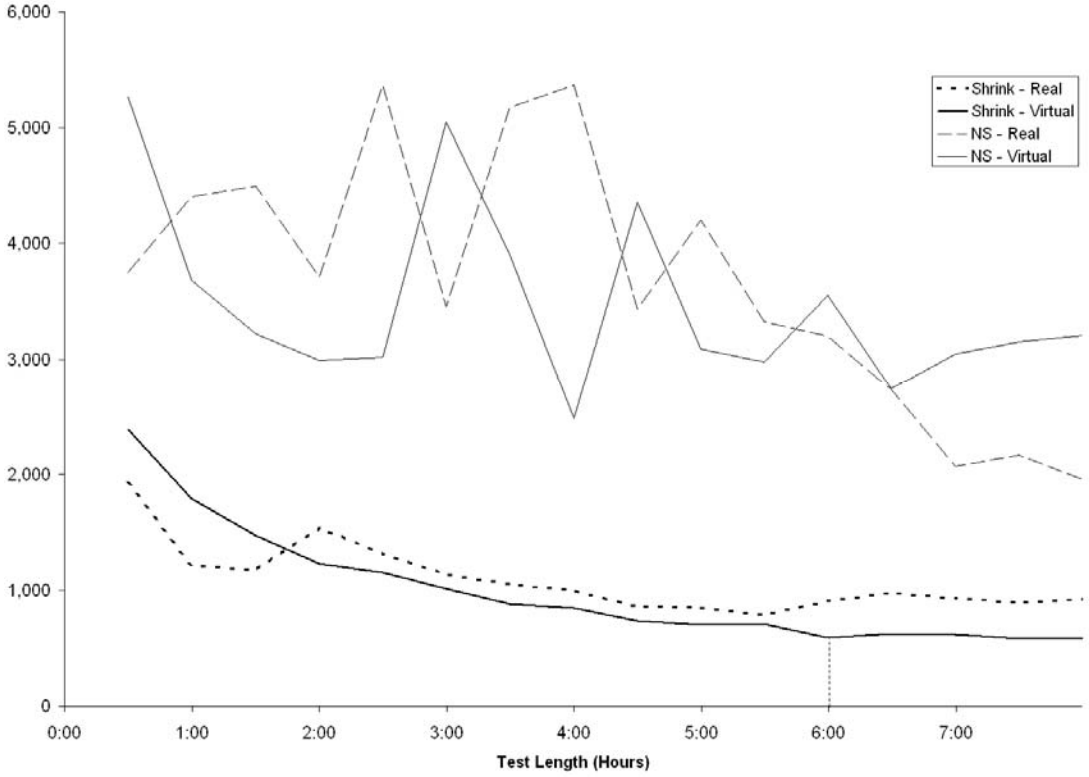


Figure 7: Comparison of models for the open split hazard parameter (δ_e^o) for a representative campaign. The dashed horizontal line represents the true value for the parameter. The Solid line represents the estimated value with a 95% confidence interval.

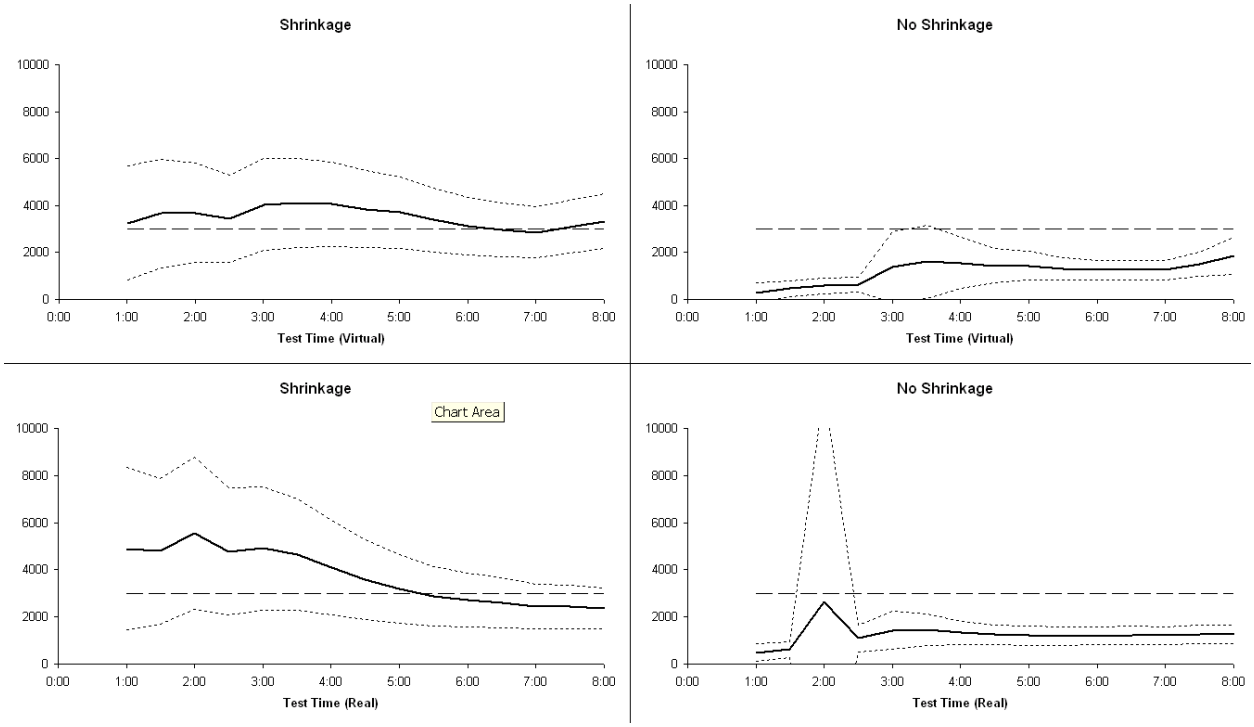


Figure 8: Mean Absolute Deviations (MAD) across models of click response. The MAD used for the click response models is made conditional on the prediction from the shrinkage-based virtual open response model. The height of the vertical dotted line represents the (14 hour) MAD for the doubling method.

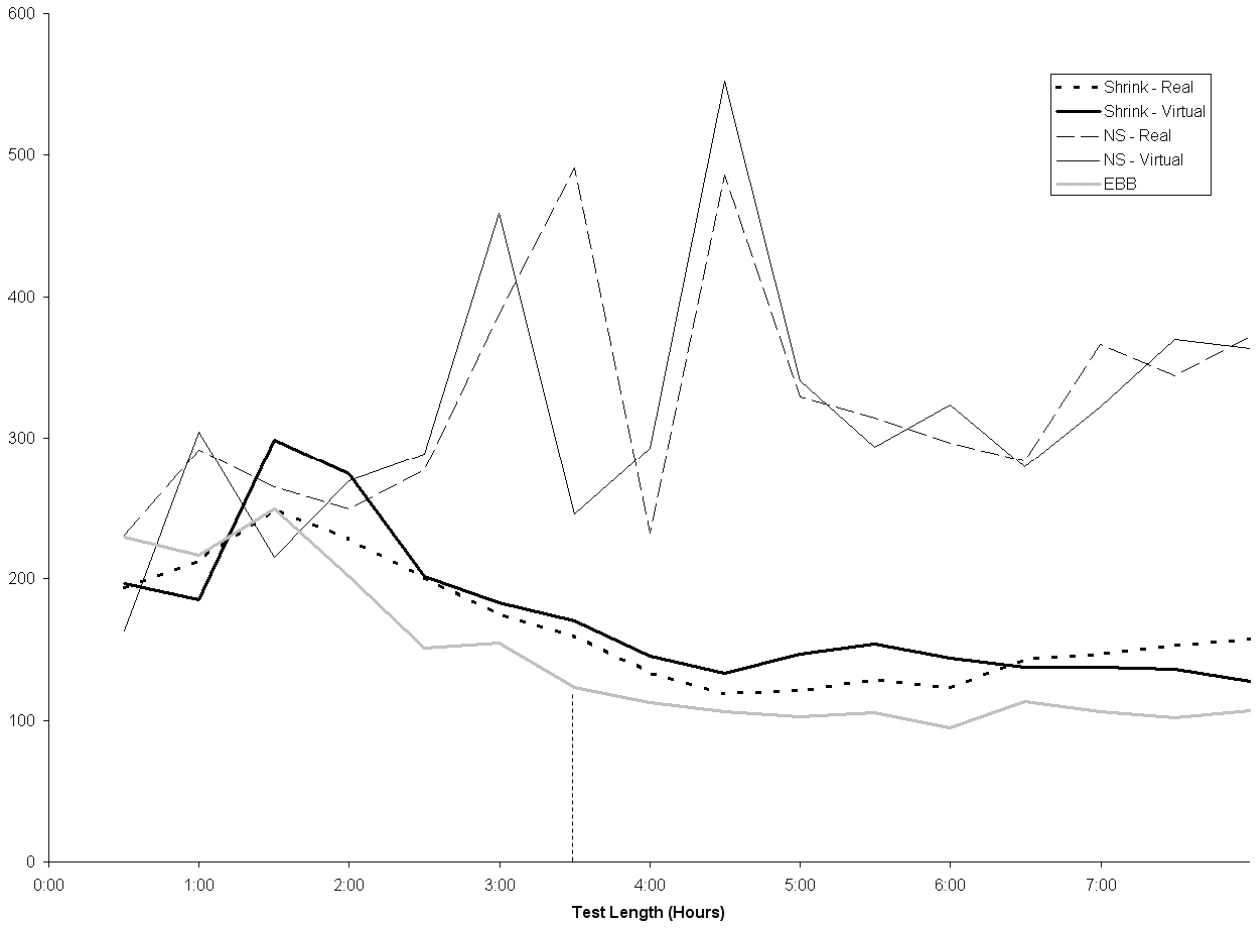


Figure 9: Test length as a function of time of day. This graph plots translates the 3.5 hour virtual waiting time into real time based on the time of day a test is performed.

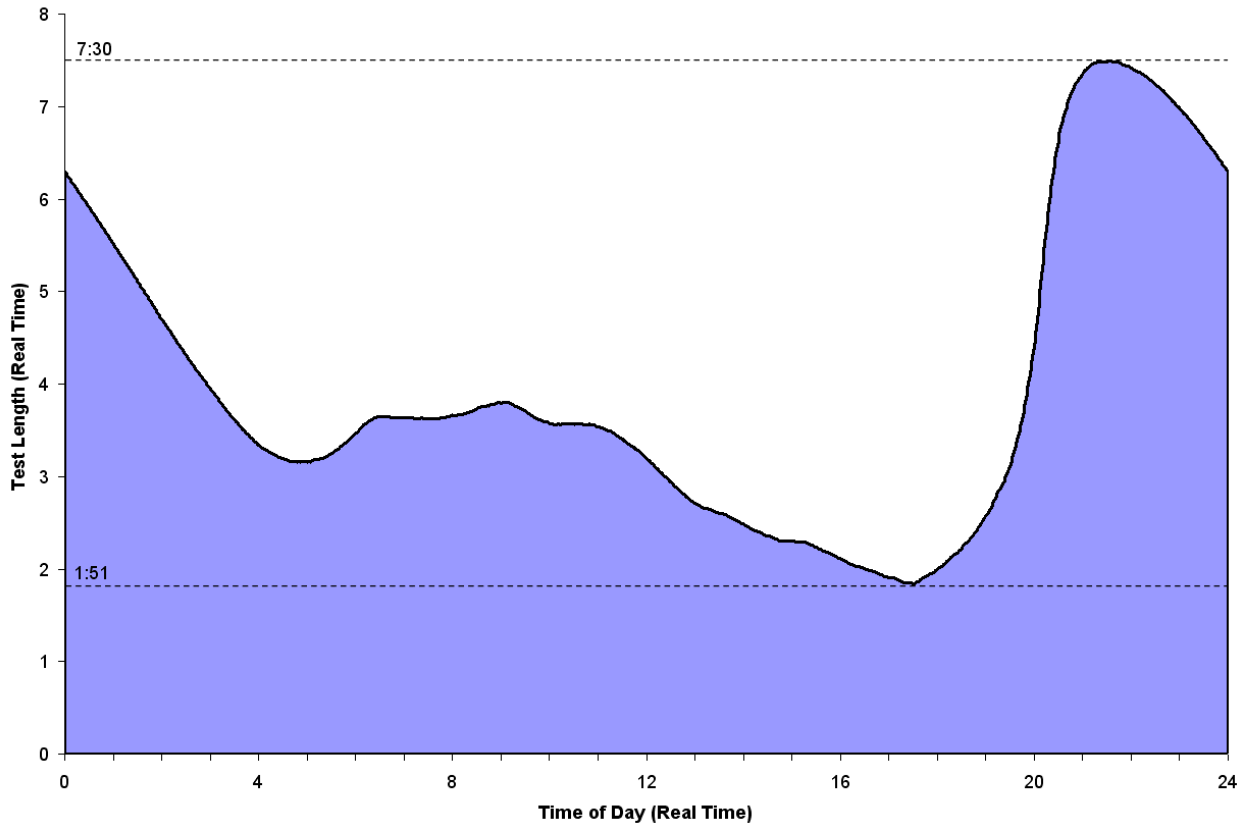


Table 1: Campaign Summary Statistics. All campaign statistics are reported based on real time.

	Mean	*Minimum*	*Maximum*	*Std Dev*
Emails Sent/Campaigns	24,681	5,171	84,465	20,393
Opens/Campaigns	4,457	600	13,116	3,877
Clicked/Campaigns	387	53	1,444	360
Open rate	0.181	0.081	0.351	0.073
Click-through | open	0.096	0.035	0.303	0.061
Click-through rate	0.016	0.005	0.105	0.021
Doubling Time (hours)	13.76	4	29	5.57
First Open (minutes)	6.41	0.07	14.68	2.08
First Click (seconds)	2.08	1.0	6.0	1.656
Number of campaigns	25			

Table 2a: Fit statistics for the virtual open time model. For the different specifications, the columns report the Log-Likelihood value (LL), the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Shaded cells show the lowest AIC and BIC for a specific campaign.

Campaign	Exponential			Weibull			Log-Normal			Log-Logistic		
	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC
1	-3150	6303	6302	-2702	5412	5408	-2684	5375	5372	-2722	5451	5448
2	-5016	10035	10034	-4142	8292	8288	-3861	7730	7726	-3868	7744	7740
3	-3699	7401	7400	-3100	6208	6204	-2971	5950	5946	-2951	5910	5906
4	-32450	64904	64902	-26424	52858	52852	-24901	49811	49806	-25015	50039	50034
5	-2869	5741	5740	-2457	4922	4918	-2352	4711	4708	-2361	4729	4726
6	-7794	15591	15590	-6418	12844	12840	-6144	12296	12292	-6163	12334	12330
7	-13780	27564	27562	-12145	24299	24294	-11289	22587	22582	-11268	22545	22540
8	-21308	42620	42618	-18126	36261	36256	-16641	33291	33286	-16613	33235	33230
9	-4089	8181	8180	-3328	6664	6660	-3112	6232	6228	-3115	6238	6234
10	-18481	36966	36964	-15523	31055	31050	-14628	29265	29260	-14540	29089	29084
11	-4056	8115	8114	-3576	7160	7156	-3462	6932	6928	-3483	6974	6970
12	-11185	22374	22372	-9793	19595	19590	-8950	17909	17904	-8918	17845	17840
13	-4938	9879	9878	-3998	8004	8000	-3865	7738	7734	-3863	7734	7730
14	-8402	16808	16806	-7126	14260	14256	-6511	13030	13026	-6473	12954	12950
15	-5842	11687	11686	-5223	10454	10450	-4954	9916	9912	-4897	9802	9798
16	-29143	58290	58288	-25437	50884	50878	-23663	47335	47330	-23651	47311	47306
17	-4269	8541	8540	-3701	7410	7406	-3576	7160	7156	-3596	7200	7196
18	-1516	3035	3034	-1244	2495	2492	-1162	2331	2328	-1155	2317	2314
19	-6747	13497	13496	-5651	11310	11306	-5209	10426	10422	-5216	10440	10436
20	-26580	53164	53162	-23524	47057	47052	-21809	43627	43622	-21847	43703	43698
21	-5038	10079	10078	-4061	8130	8126	-4006	8020	8016	-3990	7988	7984
22	-5689	11381	11380	-4997	10002	9998	-4704	9416	9412	-4688	9384	9380
23	-17916	35836	35834	-16204	32417	32412	-15250	30509	30504	-15317	30643	30638
24	-8086	16176	16174	-6735	13478	13474	-6719	13446	13442	-6698	13404	13400
25	-7140	14283	14282	-5711	11430	11426	-5700	11408	11404	-5706	11420	11416

Table 2b: Fit statistics for the virtual click time model. For the different specifications, the columns report the Log-Likelihood value (LL), the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Shaded cells show the lowest AIC and BIC for a specific campaign.

Campaign	Exponential			Weibull			Log-Normal			Log-Logistic		
	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC
1	-117	236	236	-115	235	234	-109	223	222	-110	225	224
2	-217	436	436	-207	419	418	-189	383	382	-189	383	382
3	-216	434	434	-215	435	434	-191	387	386	-189	383	382
4	-2567	5137	5136	-2506	5019	5016	-2319	4645	4642	-2316	4639	4636
5	-144	290	290	-142	289	288	-126	257	256	-124	253	252
6	-711	1425	1424	-697	1400	1398	-666	1338	1336	-672	1350	1348
7	-938	1879	1878	-908	1823	1820	-839	1685	1682	-834	1674	1672
8	-1972	3947	3946	-1921	3849	3846	-1708	3423	3420	-1669	3345	3342
9	-337	676	676	-330	666	664	-310	626	624	-310	626	624
10	-546	1095	1094	-546	1098	1096	-512	1030	1028	-516	1038	1036
11	-844	1691	1690	-842	1690	1688	-805	1616	1614	-809	1624	1622
12	-396	794	794	-386	778	776	-347	700	698	-344	694	692
13	-517	1036	1036	-506	1018	1016	-475	956	954	-475	956	954
14	-341	684	684	-340	686	684	-311	628	626	-307	620	618
15	-876	1755	1754	-876	1758	1756	-836	1678	1676	-838	1682	1680
16	-1923	3849	3848	-1887	3781	3778	-1713	3433	3430	-1693	3393	3390
17	-510	1022	1022	-486	978	976	-441	888	886	-437	880	878
18	-101	204	204	-96	197	196	-89	183	182	-89	183	182
19	-232	466	466	-221	447	446	-202	409	408	-200	405	404
20	-851	1705	1704	-840	1686	1684	-772	1550	1548	-764	1534	1532
21	-122	246	246	-118	241	240	-110	225	224	-108	221	220
22	-465	932	932	-450	906	904	-412	830	828	-407	820	818
23	-938	1879	1878	-930	1867	1864	-865	1737	1734	-859	1724	1722
24	-556	1114	1114	-537	1080	1078	-492	990	988	-486	978	976
25	-201	404	404	-199	403	402	-180	365	364	-178	361	360

Table 3: Results from the Decision Rule Simulation

	Actual	Testing Rule 1 Doubling Method at 14 Hours	Testing Rule 2 Proposed Model at 3.5 Hours	Testing Rule 3 Proposed Model at 95% CI
Number of Campaigns Selected	7	11	9	7
True Positives		7	6	5
False Positive		4	3	2
False Negative		0	1	2
Average Testing Time (Hours)		14	3.5	6.9
Minimum Testing Time		14	3.5	1
Maximum Testing Time		14	3.5	63
Click-through Rate	4.45%	3.34%	3.63%	4.00%
Improvement over No Rule	123%	67%	82%	100%
Improvement over DM	33%	0%	9%	20%