Quality, Expectations, and Customer Lifetime Value

Michael Braun
MIT Sloan School of Management
Massachusetts Institute of Technology
braunm@mit.edu

David A. Schweidel
Goizueta Business School
Emory University
dschweidel@emory.edu

Eli Stein
Harvard College
Harvard University
elistein@college.harvard.edu

March 8, 2013

Abstract

We develop a framework to estimate the impact of service quality on expected customer lifetime value (ECLV) in non-contractual settings. Based on a latent attrition model, we incorporate covariates that reflect the quality of service requested and received for each transaction. We find that service quality affects customer lifetimes asymmetrically, with the failure to meet customer expectations increasing churn more than exceeding those expectations will decrease it. However, the value of information from these quality metrics for estimating ECLV depends on the recency and frequency of customers' transaction, and is greater when there is more uncertainty about whetheror not the customer has already churned. These insights can help marketers decide how much to invest in improving or maintaining service quality, and how to allocate resources across different customer accounts. The model is easy to use, does not require simulation-based estimation methods, and is flexible enough to accomodate a general class of covariates that vary across both individuals and transactions.

According to a recent survey of chief marketing officers conducted by IBM (IBM Corporation 2011), 63% of respondents said that return on investment (ROI) would be the most important measure of success in the next three to five years. These efforts include investments in improving the quality of goods and services. Rust et al. (1995) argue in favor of assessing the return on quality, noting that investments in quality should be financially accountable. Other marketing researchers have investigated the relationship between service quality and the expectations that customers have about future interactions with the firm(Boulding et al. 1993, 1999). A separate and distinct stream of research focuses on valuation of the customer base in non-contractual (i.e., transactional) relationships. Models such as the Pareto/NBD (Schmittlein et al. 1987) and the BG/NBD (Fader et al. 2005a) leverage customer-level transaction histories to forecast future transaction counts and discounted cash flows. Although some researchers have proposed methods of incorporating time-invariant predictors into such analyses (e.g., Abe (2009) includes demographic information in his models), limited work has considered how to apply characteristics that are specific to individual transactions. How to apply transaction-specific data to assess the return on investment from marketing efforts remains an open, yet important, area of research.

In this paper, we integrate information on the quality of service encounters into a probability model of customer retention and lifetime value. Consider the following scenario. An individual goes to a local restaurant, and given some prior knowledge of the cuisine, service, atmosphere and price (either from personal experience or reputation), enters with some expectation of what that meal will be like. Suppose he has a positive experience that exceeds these expectations. Such an experience may result in the patron returning to the restaurant again at some future time. All else being equal, the level of service provided by the employees may have generated incremental revenue for the restaurant. Next, let's consider the same customer, but instead of having an experience that beat his expectations, suppose that the experience fell short. This perceived shortcoming might make it more likely that the customer stops visiting that restaurant altogether. In that case, the future revenue stream from that customer falls to zero. These examples illustrate how a single service interaction could affect the customer's future transactional activity, and consequently the revenue

that he may eventually generate for the firm. Despite this, extant empirical research on customer valuation and customer base analysis fails to consider the quality of the service interactions.

Neglecting the differences that exist across customers' service experiences omits potentially valuable information that managers can use to improve predictions of customers' future behavior. Empirical models for customer base analysis such as the Pareto/NBD (Schmittlein et al. 1987) and the BG/NBD (Fader et al. 2005a) rely on summary statistics of past transactional activity, namely the total number of transactions (frequency), and the time of the most recent transaction (recency). An advantage of using summary statistics is that the firm does not need to model each transaction separately. However, by aggregating the frequency and recency data, information about the nature of specific transactions is lost. Take the case of two customers who have engaged in the same number of transactions during the last year, and with their most recent transaction occurring on the same day. Based on the summary statistics of recency and frequency, these two customers' transaction histories are identical, so the predictions of future activity will be the same. However, if a firm had access to information on the quality of the latest service encounter relative to a customers' expectations, the firm could have different beliefs of what a customer would do in the future. Failing to meet expectations, for example, may lead the firm to believe that a customer is now at a greater risk for churn, while exceeding expectations may lead to perceptions that this risk is lower (Ho et al. 2006). Thus, the incorporation of information on the quality of service may provide firms with increased guidance for managing customer relationships compared to the information provided by transaction histories alone.

How service quality affects churn will also affect managers' expectations about if, and when, specific customers' transactions will take place. This link allows marketers to assess the long-term cash flow implications of service encounters that exceed or fall short of customer expectations. By quantifying the increased (or decreased) revenue associated with the quality of transactions, firms can decide how much to invest in service quality improvement. One interesting aspect of this decision is the value of the information that is contained in the quality metrics. For example, even if there is a large effect of service quality on churn probabilities, the incremental impact of service

quality on expected future cash flows might be quite low if the customer's baseline tendency to churn is already high (Braun and Schweidel 2011).

In the next section, we provide a brief review of the literature on service quality. Afterwards, we describe a parsimonious modeling framework, based on the models often employed in the customer relationship management and customer valuation literature, that incorporates service quality into a latent attrition modeling framework. We then present an empirical analysis, using a dataset from a noncontractual service provider, and demonstrate how to use our framework to quantify the impact of service quality on expected customer lifetime value.

1 Related Literature

Several marketing researchers have studied service quality and its relationship to customer expectations. Boulding et al. (1993) find that a customer's evaluation of a service encounter is affected by his prior expectations of what *will* and *should* occur, as well as the quality of service delivered on recent service encounters. In essence, *would* and *should* expectations for a service encounter are a weighted average of prior expectations and the recently experienced service. Boulding et al. (1999) further investigate the process by which expectations are updated. In addition to affecting a customer's cumulative opinion, the authors find that prior beliefs also affect how experiences are viewed. As a result, prior expectations deliver a "double whammy" to evaluations of quality. This suggests that service encounters are not all equal in the eyes of consumers, as the way in which service encounters are viewed are affected by past experiences. For example, in our restaurant example, the exact same level of quality might exceed expectations in a mid-range family restaurant, but miss expectations in a fancy bistro. Yet, extant customer valuation models in both non-contractual and contractual settings often assume that the "touch points" associated with customer-firm transactions are equivalent to each other (Fader et al. 2005a; Schweidel et al. 2008). In this research, we seek to incorporate the finding from the service quality literature that characteristics specific to a service encounter affect the likelihood that customers remain in active relationships with the service

provider.

Rust et al. (1999) further investigate the role of customer expectations in perceptions of quality. Rather than focusing on the average, the authors highlight the importance of the distribution of customer expectations. The authors tackle a number of myths that had been held in regards to the level of service that providers should deliver to their customers. In contrast to the popularly held belief that exceeding expectations is necessary, the authors find evidence to suggest that simply meeting customers' expectations can result in a positive shift in preferences. The authors also find that service encounters that are slightly below expectations may not affect customers' preferences. Though provoking, the authors recognize that because they conducted their investigation in a laboratory setting and relied on self reports, there is a need for additional research on these effects.

In addition to the work that has been conducted on service quality, our work is also related to research conducted into customer satisfaction. Bolton (1998) investigates the impact of customer satisfaction on the duration for which customers continue to subscribe to a contractual service. She finds that reported customer satisfaction with the service, solicited prior to the decision of whether to remain a subscriber or cancel service, is positively related to the duration for which a customer will retain service. She also finds evidence to suggest that recent experiences with the service provider are weighed differently depending on whether the experience was evaluated as positive or negative.

While Bolton (1998) examines the link between satisfaction and service retention in a contractual setting, Ho et al. (2006) investigate this relationship in a non-contractual setting using an analytical modeling framework. Building upon the Pareto/NBD model (Schmittlein et al. 1987), they show that there may exist an optimal level of investment in customer satisfaction depending on the costs associated with increasing customer satisfaction. The authors suggest augmenting the RFM scoring approach commonly used in direct marketing activities with and RFMS scoring approach that accounts for recency, frequency, monetary value and satisfaction.

Using customer transaction and quality data from a B2B service provider, we empirically investigate the impact of customers' transactional experiences on future behavior in a non-contractual environment. Consistent with Bolton (1998), we allow for exceeding the quality of service that a

customer requests and for failing to meet the quality level requested to have differing effects. We quantify the effects of meeting customers' quality expectations, exceeding expectations and falling short of expectations by examining how such experiences impact future customer transactional behavior (Rust and Zahorik 1993). In addition to assessing the impact of exceeding or falling short of customer expectations empirically, we also demonstrate the value of quality assessments by examining the extent to which expectations of customers' future behavior are affected by transactions in which expectations were met, exceeded or not met.

2 Model

In this section we propose a general form of a latent attrition model that incorporates transaction-specific effects. To keep terminology consistent with the empirical example we will describe later in this paper, we say that the client of the firm places orders for jobs, and the firm fills those orders by completing the jobs. Thus, orders and jobs always occur in a pair, and are indexed by k. We assume that these jobs are completed the instant the order is placed, so we index calendar time for orders and jobs by t. Without loss of generality, we define a unit of calendar time as one week. The service was introduced to the marketplace at time $t = 0$ and T is the week of the end of the observation period. Let t_1 be the week of the client's first order, let x be the number of orders between times t_1 and T, *including* that first order at t_1, and let t_k be the time of order k. Therefore, t_x is order time of the final, observed job. For clarity, we are suppressing the client-specific indices on t and x in the model exposition.

Our baseline model is a variant of the BG/NBD model for non-contractual customer base analysis (Fader et al. 2005a). Immediately before the client places an initial order at time t_1, he is in an "alive" state. While alive, the client places orders according to a Poisson process with rate λ. After each job (including the first one), a customer may churn, resulting in that order being his last. With probability p_k, the customer churns after the k^{th} job and transitions from the "alive" state to the "death" state. Upon doing so, we assume that the customer is lost for good and will not place

any more orders, ever. If the customer does not churn, then the time until the next order, $t_{k+1} - t_k$, is a realization of an exponential random variable with rate λ. We never observe directly when, or if, a client churns, although if a client places x orders, he must have survived $x-1$ possible churn opportunities.

For a client who places x orders between times t_1 and T, the joint density of the $x-1$ inter-order times is the product of $x-1$ exponential densities. For this client, there could not have been any orders between times t_x and T. This "hiatus" could occur in one of two ways. One possibility is that the client may have churned after job x, with probability p_x. Alternatively, the client may have "survived" with probability $1-p_x$, but the time of the next order would be sometime after T. Thus, conditional on surviving x jobs, the probability of not observing any more jobs before time T is $e^{-\lambda(T-t_x)}$. Hence, the conditional data likelihood for a single client is

$$f(x, t_{2:x}|\lambda, p_{1:x}) = \lambda^{x-1} e^{-\lambda(t_x - t_1)} \left[\prod_{k=1}^{x-1} (1-p_k)\right] \left[p_x + (1-p_x) e^{-\lambda(T-t_x)}\right] \tag{1}$$

Next, let q_k be a non-negative function that can influence the churn probability after job k, such as a vector of covariates that varies from job to job. Define the probability of churning after job k as $p_k = 1 - e^{-\theta q_k}$ and define $B_k = \sum_{j=1}^{k} q_k$. Substituting these definitions into Equation 1,

$$f(x, t_{2:x}|\lambda, \theta, q_{1:x}) = \lambda^{x-1} e^{-\lambda(t_x - t_1) - \theta B_{x-1}} \left[1 - e^{-\theta q_x} + e^{-\theta q_x} (1 - \lambda (T - t_x))\right] \tag{2}$$

This expression of the likelihood assumes that all clients place orders at the same rate, and that all clients have the same baseline propensity to churn after each job. To incorporate heterogeneity of latent characteristics into the model, we let λ and θ vary across the population according to gamma distributions, where $\lambda \sim \mathcal{G}_\lambda(r, a)$ and $\theta \sim \mathcal{G}_\theta(s, b)$. Integrating out these latent parameters, we get the marginal likelihood:

$$\mathcal{L} = \frac{1}{P(\mathcal{A})} \frac{\Gamma(r+x-1)}{\Gamma(r)} \left(\frac{b}{b+B_x}\right)^s \left(\frac{a}{a+T-t_1}\right)^r \left(\frac{1}{a+T-t_1}\right)^{x-1} \tag{3}$$

where

$$P(\mathcal{A}) = \left[1 + \left(\frac{a+T-t_1}{a+t_x-t_1}\right)^{r+x-1}\left[\left(\frac{b+B_x}{b+B_{x-1}}\right)^{s} - 1\right]\right]^{-1} \tag{4}$$

In Appendix A.2, we show that $P(\mathcal{A})$ is equal to the probability that a client with observed data x, t_1, t_x, B_{x-1} and B_x, has not yet churned by time T.

An attractive feature of this model is that one can estimate the parameters using maximum likelihood, without resorting to simulation-based approaches. To do this, one would simply compute the sum of the logs of Equation 3 for each client, and maximize with respect to r, a, s, b, and any parameters in the terms that comprise B_x. However, to maintain this important degree of computational efficiency, there are some tradeoffs. Although we do allow for unobserved heterogeneity in λ and θ by carefully choosing mixing distributions that generate a closed-form marginal likelihood, we assume that the parameters that determine B_x are homogeneous. As the churn process is latent and occurs at most once for each customer, even if we could derive an efficient algorithm to get individual-level estimates of these parameters, they may be difficult to identify and may be highly sensitive to the assumed prior specification. This is because the elements of q_k are influencing the *unobserved* quantity p_k. While there is enough information in the dataset to estimate population-level parameters, the individual-level estimates may be unreliable.

Also, we assume that the mixing distributions of λ and θ are independent. Whether one could effectively estimate any degree of correlation between latent parameters in any latent attrition model remains an unsettled topic. Fader et al. (2010) show that for the BG/BB (the discrete time analog of the BG/NBD), the incorporation of correlation between the transaction process and the latent attrition process does not yield substantively different findings. Abe (2009) allows for latent correlation in a model quite similar to the Pareto/NBD, and found no significant correlation in any of his three datasets. Furthermore, as shown in Appendix A.1, it is clear that the individual posterior densities will be correlated, even if the priors are not; a similar result is in Fader et al. (2010). Thus, we believe that it is worthwhile to keep the model sufficiently simple that we can

retain the computational advantages from a closed-form likelihood function, and hence the practical value of the model.

2.1 Conditional expectations and ECLV

Once a manager has parameter estimates in hand, he might be interested in the number of orders that we might receive from a newly acquired client during a period of t weeks. In Appendix A.3 we show that the *prior* expected value of this order count is

$$E[X(t)] = \sum_{k=1}^{\infty} \left(\frac{b}{b+B_k}\right)^s \widetilde{\mathbb{B}}\left(\frac{t}{a+t}; k, r\right) \tag{5}$$

This manager might also want to know how many orders he can expect from an existing customer, during the next t^* periods, given an observed transaction history. In Appendix A.4, we show that this *posterior* expected value is

$$E^*[X(t^*)|x,t_x,B_x] = P(\mathcal{A}) \times \sum_{k=1}^{\infty} \left(\frac{B_x+b}{B_x+B_k+b}\right)^s \widetilde{\mathbb{B}}\left(\frac{t^*}{t^*+a+T-t_1}; k, r+x-1\right) \tag{6}$$

The function $\widetilde{\mathbb{B}}(x;a,b)$ is the cdf of a beta distribution, with parameters a and b, evaluated at x. [1] (a glossary of many of the functions we use in this paper is in Table 5 in Appendix A). We also include in Appendix A.5 the prior and posterior probability mass functions for the number of orders (i.e., to express the probability of placing a particular number of orders during some future number of weeks).

Next, suppose that it is now time T, and that we observe the order history for a single client. Although it is useful to know the expected number of future orders, the wise manager will recognize that those orders will come at different times. Given the time value of money, orders that occur soon are more valuable than orders that are placed later. Without loss of generality, we will assume that the company accrues a profit of one dollar for each job. Let δ be a discount factor that captures the time value of money, so a dollar earned t weeks from now is worth δ^t today (for notational

[1] In Equation 6, index of summation k is "reset", so it refers to potential orders that are made after time T.

simplicity, we reset the counter of t so $t = 0$ at T, and we assume that payments are made at the end of the week). The expected CLV (ECLV) for this client is the sum of discounted incremental expected orders.

$$\begin{aligned} ECLV &= \sum_{t=1}^{\infty} \left(E^*[X(t)|x,t_x,B_x] - E^*[X(t-1)|x,t_x,B_x] \right) \delta^t \\ &= P(\mathcal{A}) \sum_{k=1}^{\infty} \left(\frac{B_x+b}{B_x+B_k+b} \right)^s \\ &\quad \times \sum_{t=1}^{\infty} \delta^t \left[\widetilde{\mathbb{B}} \left(\frac{t}{a+T-t_1+t}; k, r+x-1 \right) - \widetilde{\mathbb{B}} \left(\frac{t-1}{a+T-t_1+t-1}; k, r+x-1 \right) \right] \\ &= (1-\delta) P(\mathcal{A}) \sum_{k=1}^{\infty} \left(\frac{B_x+b}{B_x+B_k+b} \right)^s \sum_{t=1}^{\infty} \delta^t \widetilde{\mathbb{B}} \left(\frac{t}{a+T-t_1+t}; k, r+x-1 \right) \end{aligned}$$

These future cash flows depend on a number of different elements. The parameters r, a, s and b capture the distribution of order rates and baseline churn likelihoods across the population. Clients with low t_x and high x might be more likely to have already churned, so there is a low probability of receiving any cash flows from them in the future. Clients with high x and high t_x are more likely to be alive, and to order often, so their ECLV should be high.

Note that the expressions for $E[X(t)]$, $E^*[X(t^*)|x,t_x,B_x]$ and $ECLV$, include an infinite sum across jobs, indexed by k. Also, recall that $B_k = \sum_1^k q_k$. If q_k is a function of job-specific covariates, then we would never actually *observe* q_k for any $k > x$. We approximate B_k by replacing it with its expected value. As B_k is linear in q_k, we can get the expected value of B_k through the expected values of q_k. Thus, one way to approximate the conditional expectations of x and $ECLV$ would be to implement a probability model for q_k. We will provide an example with our empirical application. Without this heuristic, the only way to compute these conditional expectations would be to numerically integrate them over all possible values of B_k, such as by using simulation.

Like all statistical models, this model is intended as a schematic of the actual data-generating process. To give the model some useful parametric structure, we treat the latent attrition process as a manifestation of a random variable. Though one can always propose more complicated versions of a model, such as allowing for duration dependence in purchase times or contagion across clients in

their propensities to churn, we favor parsimony so as to avoid overparameterizing the model given certain limitations in the dataset.

3 Empirical Analysis

The context in which we study the role of quality on customer lifetime value is that of an online market for freelance writing services. The firm in question operates a website on which clients can post specifications for "jobs," and writers can claim jobs to complete. An example of a job is a 100-word product description on an online retailer's website. Another is a 500-word summary of what participants at a conference might do for fun when exploring the host city. Job specifications include all of the information a writer would need to complete the job: the topic area (e.g., sports, health) , intended audience, word count, and so forth. Clients are encouraged to be as specific as possible in their requirements, as that makes it more likely the client will be satisfied with the results. In our taxonomy, we consider an "order" to be equivalent to the posting of a job.

Clients also choose a minimum quality rating for the writers who are eligible to claim the projects. The firm maintains a bank of reviewers to screen and rate the writers who are eligible to claim jobs. These reviewers are employed directly by the firm, and are considered to be experts in writing quality (most have Masters of Arts degrees, or similar qualifications). Upon initial application, a writer submits a writing sample, and a reviewer rates the writer as A, B, C or D. The firm's website provides examples of work from the different rating categories, so clients have a general idea about the differences to expect across the different ratings. Ratings differ according to objective criteria such as accuracy, grammar, style and vocabulary. A D-rated writer might produce work with spelling errors and simple sentence structure with no creative insight, while work from an A-level writer will be of professional quality.

Clients pay, and writers earn, on a per-word basis, where the charge for each word depends on the quality rating of the *job*, and the number of words in the specification. By selecting a quality rating, the client knows exactly how much he will pay per word. Writers claim jobs from a list of

available specifications, and jobs are assigned on a first-claim basis, so there is no bidding involved. Writers are able to claim jobs that are rated below their own ratings (e.g., an A-rated writer can choose a project from any level, but a B-rated writer cannot choose an A-rated project). In such cases, though, writers are paid the lower per-word fee. Thus, it is in the writer's best financial interest to claim jobs that match his quality rating if such jobs are available. Since it is also in the writer's best interest to complete jobs quickly, he might choose to accept jobs that are within his area of expertise. For example, a sports fan might select a job for a blog entry on a football website, while someone who likes movies might prefer to write film reviews. The company claims that they have not experienced shortages of writers, such that they could not meet client demand; most job specifications are claimed with a day, and writers have another day to complete the job. However, according to the company, most jobs are completed within 24 hours of posting.

Sometime after the writer returns the completed job to the client, the firm's bank of reviewers assigns each job a grade. Clients are not involved in this rating process, and neither clients nor writers ever see the grade for a particular job. However, a writer's rating can be adjusted according to his rating history. This gives the writer an incentive to complete the job well; if his jobs are rated poorly, he will be downgraded and no longer eligible for higher-paying jobs. The reviewers try to rate jobs as accurately and objectively as possible, as a way to ensure that clients receive they quality they pay for. Writers can only be elevated to the A level manually, so the firm classifies all A-rated and B-rated jobs together in an A/B class. Reviewers may also assign a grade of E for completed jobs that do not meet even minimum standards.

3.1 Data summary

Our master dataset includes *all* completed jobs from the launch of the company in June 2008 to the end of our observation period at the end of July 2011. We are restricting our analysis to clients who are located in either the United States or Canada, for which the language of the job is English, and whose first order takes place before the end of 2010. This dataset contains information on 24,059 completed jobs that were ordered by 3,048 distinct clients. For each job, we have identifiers for

the client and writer, the day that the order was placed, some other details of the job specification. We also have the requested rating of the job, as well as the grade the job received from the bank of reviewers. Table 1 shows the number of jobs requested at each quality rating, and the quality grade of the work that the writer delivered to the client. By exploiting variation in the ratings, we can examine the impact of quality level delivered (assessed objectively by the reviewer), relative to the level that was requested by the client, on customers' future transactional activity.

		Post-hoc quality grade				
		A/B	C	D	E	Total
	A	769	0	0	0	778
Requested	B	8764	614	5	1	9384
Rating	C	2035	5733	255	18	8041
	D	1685	3280	803	88	5856
	Total	13253	9636	1063	107	24059

Table 1: Number of jobs requested at each quality rating, and the quality grade of the work that the writer delivered to the client.

We define the observation period for a particular client as the time from the day of a client's first order (t_1), until the end of our observation period (T). This is consistent with extant research in customer base analysis (Schmittlein et al. 1987; Fader and Hardie 2001; Fader et al. 2005a). If a client places x orders during that observation period, his "frequency" is equal to $x/(T-t_1)$. A client's "recency" is t_x, the week of the most recently observed order. Each day is represented as 1/7 of a week.

We recognize that some of the firm's "early adopters" might behave differently than those customers whose first order came later. In our subsequent analysis, we divide the client base into four cohorts, based on the week of the first order. Table 2 summarizes the average the number of orders, frequency and recency for each of these four cohorts.

	Num	Start	End	Average across clients				Num repeaters	
Cohort	Clients	Week	Week	$T-t_1$	t_x	freq calib	freq hold	calib	hold
0	588	1	33	116.3	29.9	.06	.01	310	552
1	568	33	66	79.1	69.1	.12	.03	462	494
2	911	66	99	46.3	94.7	.16	.03	726	750
3	981	99	130	14.0	119.6	.58	.08	678	647

Table 2: Summary statistics by cohort. All times are in weeks.

3.2 Modeling quality

Within the context of the basic model, covariates that vary from job to job are incorporated through the specification of q_k. For this dataset, we will consider models of the form

$$\log q_k = \omega \log q_{k-1} + (1-\omega)\beta' z_k \tag{7}$$

where z_k is a vector of job-specific covariates, β is a vector of coefficients, and ω is a "decay" parameter that can take values between zero (no persistance of quality stock from job to job) to one (no effect of the job-specific experience). The z_k vectors also include some invariant indicators to capture differences across cohorts (with Cohort 1 as the base level), and whether a job is the first one that a client has ordered.

To assess the impact of quality on repeat transactional behavior, we consider a set of models with different specifications. Model 1 is our baseline model, with no job-specific characteristics. As such, the role of quality is not considered. In Model 2, we include covariates that capture the main effects of the requested quality levels. Under this model, we allow for the possibility that a client who requests a Grade A job may be more (or less) likely to return and place another order than a client who requests a Grade B job. As with Model 1, Model 2 does not consider the quality that is delivered. Model 3 includes all of the effects from Model 2, as well as the main effects related to the quality of the writing that was actually delivered. The quality effects are categorized as being either the same level as requested (S), lower than requested (L) or higher than requested (H). Model 4 includes the main effects from Model 3, plus interactions. For example, exceeding the requested quality level by delivering an B grade job when a C grade job was requested may have a weaker

effect than exceeding the requested quality of a D grade job by delivering a C grade job. Finally, Model 5 allows for accumulative effects of meeting, falling short or exceeding expectations. We operationalize this by considering the quality delivered, relative to the requested grade level, on the current job and the immediately prior job. Note that Model 1 does not include a "first job" dummy, and Model 5 splits the "first job" effect across multiple dummies, depending on how the job is rated.

We summarize the models we estimate in Table 3. The parameters for each model include r, a, s, b, the three parameters for the cohort effects, the "first job" parameter(s), the decay parameter (when present), and the elements in β.

Model	Description	Num Params
1	No effects for job characteristics	8
2	Dummy variables for main effects of the requested grade rating (baseline level is Grade B)	12
3	Main effects from Model 2, plus main effects for the evaluated rating class. Dummy variables indicate of the quality of the job is lower (L), the same (S), or higher (H) that the class that was ordered. The baseline level is S.	14
4	The main effects from Models 2 and 3, plus a full set of interactions.	20
5	Main effects from Model 2, plus dummy variables that indicate the combination of evaluated ratings from the current and previous jobs. For the first ordered job, the "previous" job is indicated as F (e.g., FL, FS and FH). For subsequent jobs, possible levels are LL, LS, LH, SL, SS, SH, HL, HS, and HH. The baseline level is SS. This model does not include a special "first job" spike.	22

Table 3: Model specifications for $\log q_k$

While there are a myriad of alternative model specifications for $\log q_k$, the set presented in Table 3 allows us to consider different aspects of how quality may relate to customer retention that have been identified in prior research and assess their contribution to the performance of the model. Among the members of this set of candidate models, Models 1 and 2 do not incorporate any quality measures. Models 3, 4, and 5 allow the quality of the job, relative to the quality level that the client question, to affect the probability of customer churn, in different fashions. Including the decay parameter is consistent with research on service quality (Boulding et al. 1993), and we found that all models that included the decay parameter fit better than their decay-free counterparts. In

the interest of parsimony, we are not modeling the latent *rates* as functions of quality, nor are we modeling the decision around which level of quality to order.

3.3 Parameter estimates and model assessment

Maximum likelihood estimates of the parameters, along with 90% confidence intervals, are in Table 4. The parameters that we care most about are `rat.lower` and `rat.higher` in Models 3, 4 and 5. These are the main effects of whether a job is rated lower or higher than what was ordered (the baseline case is that the job was rated at the same quality level that the client ordered). As we would expect, if a client receives a job that is rated lower than what was was ordered, there is a statistically significant increase in the probability of churn immediately after that job. However, we do not see a significant effect when the job is rated higher than what was ordered. This asymmetry in the effect of quality is consistent with the idea that losses loom larger than gains (Kahneman and Tversky 1979; Hardie et al. 1993). Our findings are also in line with prior research by Bolton (1998), who found that perceived losses adversely impact the duration of a customer's relationship in a contractual setting while perceived gains did not have a significant impact on the duration of the relationship.

In Appendix B, we compare the relative fits of these models using a combination of different test statistics. When comparing relative model fit, we care about how well a model predicts for both a sample of clients that was used to calibrate the model, as well as for alternative sets of clients that were held out of the model estimation process. In addition, we looked at not only how well a model fits data from the 130-week calibration period, but also how well it predicts future behavior, at both the aggregate and individual levels. All of the tests in Appendix B suggest that models that incorporate quality metrics fit better than those that do not. In some cases there is disagreement about which of Models 3, 4 or 5 we should prefer. Although many of the test statistics of model fit for these models are similar, there is sufficient evidence in favor of Model 5 that we will use that model for the subsequent analysis. [2]

[2]Detailed findings from the alternative model specifications are available from the authors upon request.

	Model 1			Model 2			Model 3			Model 4			Model 5		
param	low	**pars**	high	low	**pars**	high	low	**pars**	high	low	**pars**	high	low	**pars**	high
r	.83	**.89**	.95	.84	**.90**	.96	.84	**.90**	.96	.84	**.90**	.96	.84	**.90**	.96
a	.69	**.75**	.82	.70	**.77**	.83	.70	**.77**	.83	.70	**.77**	.83	.70	**.76**	.83
s	.96	**1.07**	1.18	.97	**1.09**	1.20	.97	**1.09**	1.20	.98	**1.09**	1.20	.84	**.98**	1.12
b	1.02	**1.37**	1.72	.79	**1.09**	1.40	.80	**1.12**	1.43	.84	**1.16**	1.49	.41	**.75**	1.10
omega				.01	**.25**	.48	-.17	**.10**	.37	-.15	**.09**	.33	.22	**.41**	.60
first.job	-.28	**-.17**	-.06	-.34	**-.23**	-.12	-.33	**-.22**	-.11	-.33	**-.22**	-.11			
cohort.2	-1.30	**-1.11**	-.93	-1.35	**-1.16**	-.97	-1.35	**-1.16**	-.97	-1.35	**-1.16**	-.97	-1.41	**-1.20**	-.99
cohort.3	-1.12	**-.95**	-.78	-1.19	**-1.01**	-.84	-1.18	**-1.01**	-.83	-1.17	**-1.00**	-.83	-1.25	**-1.06**	-.86
cohort.4	-.91	**-.73**	-.56	-1.00	**-.82**	-.64	-1.00	**-.82**	-.64	-.98	**-.80**	-.62	-1.07	**-.87**	-.67
order.A				.72	**.92**	1.13	.67	**.87**	1.07	.71	**.91**	1.11	.45	**.79**	1.13
order.C				-.45	**-.33**	-.21	-.42	**-.30**	-.18	-.41	**-.28**	-.16	-.50	**-.36**	-.22
order.D				-.69	**-.54**	-.40	-.68	**-.51**	-.33	-.66	**-.40**	-.14	-.64	**-.43**	-.22
rat.lower							-.00	**.18**	.37	.18	**.41**	.63	.07	**.49**	.91
rat.higher							-.18	**-.03**	.11	-.07	**.09**	.24	-.17	**.11**	.38
prev.rat.first													-.73	**-.51**	-.28
prev.rat.lower													-.58	**-.17**	.23
prev.rat.higher													-.44	**-.17**	.09
consec.rat.FL													-.85	**-.36**	.14
consec.rat.FH													-.30	**.01**	.32
consec.rat.LL													-.65	**.00**	.66
consec.rat.LH													-1.17	**-.26**	.64
consec.rat.HL													-1.02	**-.22**	.58
consec.rat.HH													-.91	**-.51**	-.12
inter.D.lower										-.35	**.29**	.92			
inter.D.higher										-.52	**-.21**	.10			
inter.C.lower										-1.05	**-.60**	-.15			

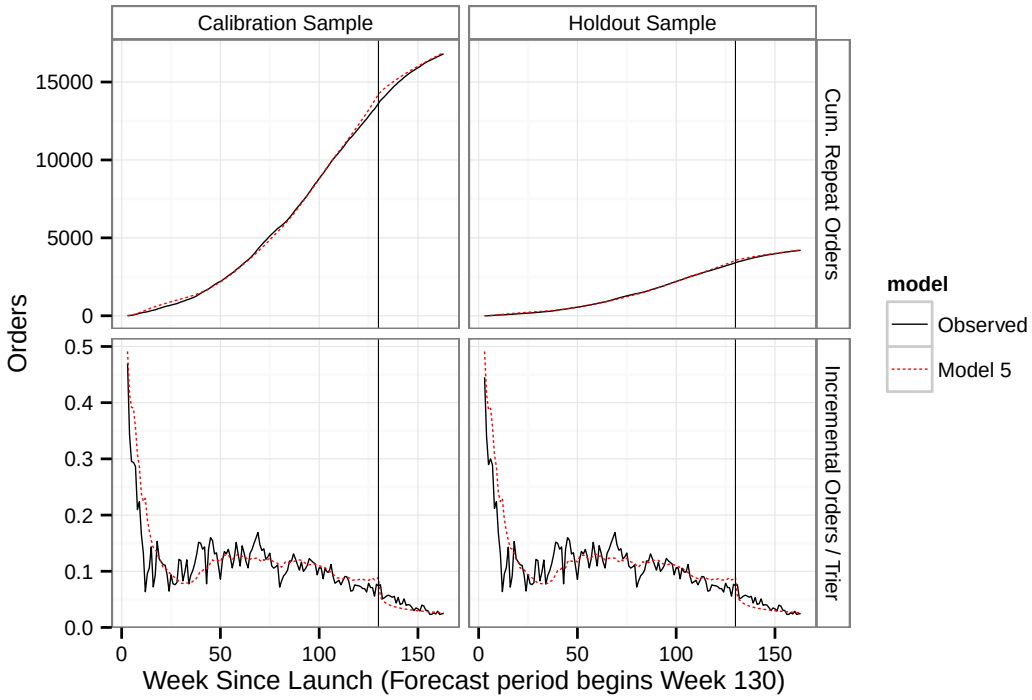
Table 4: Parameter estimates, with upper and lower bounds of the 90 percent confidence interval.

While the results in Appendix B provide measures of relative model peformance, they do no shed light on how well the model captures the observed behavior. As further assessments of the performance of Model 5, Figure 1 illustrate model fit at the aggregate level. Figure 1a plots the cumulative and incremental number of weekly orders, for both in-sample and holdout populations. The vertical lines at Week 130 divide the calibration and forecast time periods. We used only data to the left of the lines for estimating the model parameters, and we included only those clients whose initial order was before Week 130. Here, we see that Model 5 does well in tracking the number of orders from week to week. In Figure 1b, we compare the histogram of pre-client order counts with the distribution of counts from Model 5 predicts. Again, Model 5 appears to fit rather well.

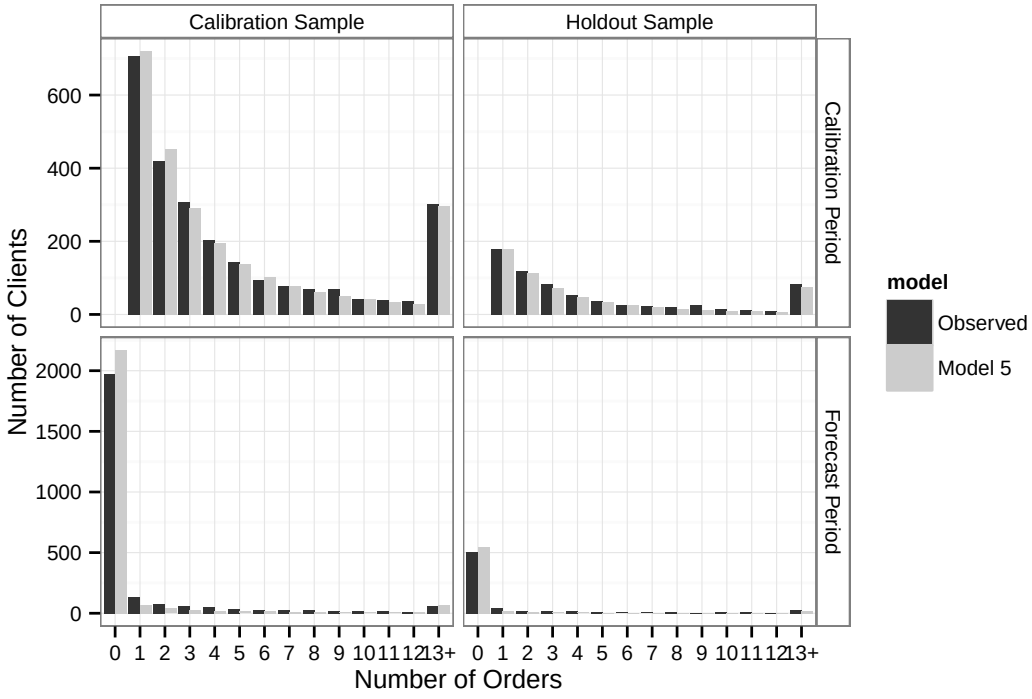
At the individual level, one statistic of managerial interest is $P^*(0) = P(X^*(t^*) = 0|x, t_x, \cdot)$, the posterior probability that a client will place no orders during the forecast period. We tested how well Model 5 predicts which clients will order during the forecast period by examining the frequentist properties of the estimate. First, we assigned each client to one of 15 "bins", according to the client's posterior $P^*(0)$. A client is assigned to bin i if $(i-1)/15 \ < \ P^*(0) \ \leq \ i/15$ for $i = 1 \ldots 15$. Next, we computed the observed proportion of clients in each bin who did not place an order during the forecast period. We consider the model to be well-calibrated if the predicted probabilities and observed proportions are aligned. Figure 2 confirms that they are. Each dot represents the membership of the bin. The x-coordinate is the midpoint of the bin, and the y-coordinate is the observed incidence of "no orders" for the members of that bin. "Perfect" calibration would have occurred if all of the dots fell exactly on the 45° line. Of course, we would expect some random variation around this line, so we can still be confident that Model 5 forecasts the incidence of future orders, at the client level, quite well.

4 Effect of quality on ECLV

Using Model 5 as our "model of choice," we can now examine how knowledge of the quality of a job affects our predictions of ECLV. There are two different questions we can ask.



(a) Weekly incremental repeat orders per previously acquired customer. The vertical line is at Week 130, the end of the calibration period and the start of the forecast period.



(b) Observed and predicted histograms of orders.

Figure 1: Fit and forecast assessment for Model 5.

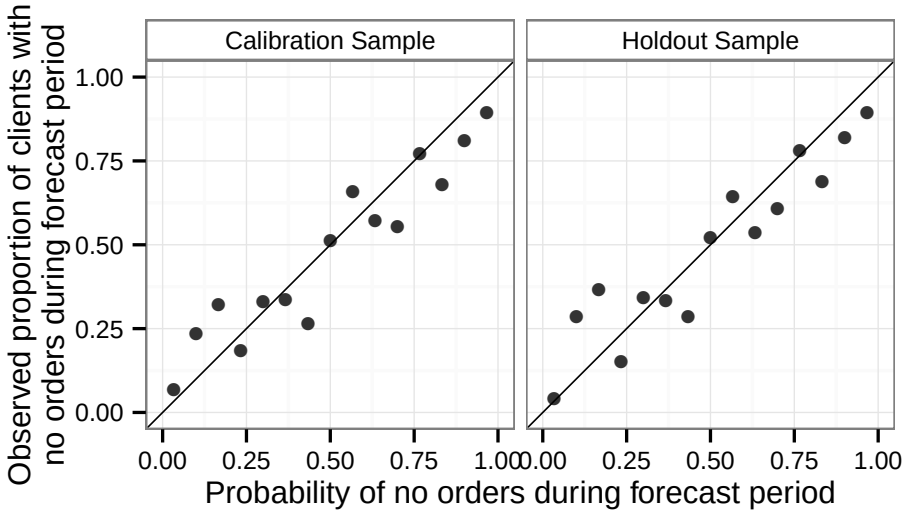


Figure 2: Predicted vs observed probability of a particular client making no orders during the forecast period.

1. At the moment immediately following the completion of a job, what is the effect on ECLV if the client receives a different level of quality from what the client ordered?

2. At time T, what is the effect on ECLV if the level of quality of the most recent job was different from what the client ordered?

Before answering these questions, we need to consider how ECLV depends on frequency (expressed as x) and recency (t_x). Figure 3 plots the contours that connect the same levels of ECLV, at $T = 130$, for hypothetical clients who placed the first order at time $t_1 = 1$, who requested quality grades B, C or D for the most recent order, and whose last order was rated lower, the same, or higher than what was ordered (to keep the number of possible combinations manageable, we assume that all other jobs were rated as ordered). These kinds of iso-value curves were introduced by Fader et al. (2005b) for the Pareto/NBD model, to illustrate how ECLV depends on x and t_x. Since ECLV depends on the probability that a customer remains active, we expect to see these backward-bending contours(Fader et al. 2005b). When the number of orders is large and the most recent order was in the distant past, it is more likely that the client has already churned, compared to the scenario in which the same number of orders was made but the x^{th} order was made more recently.

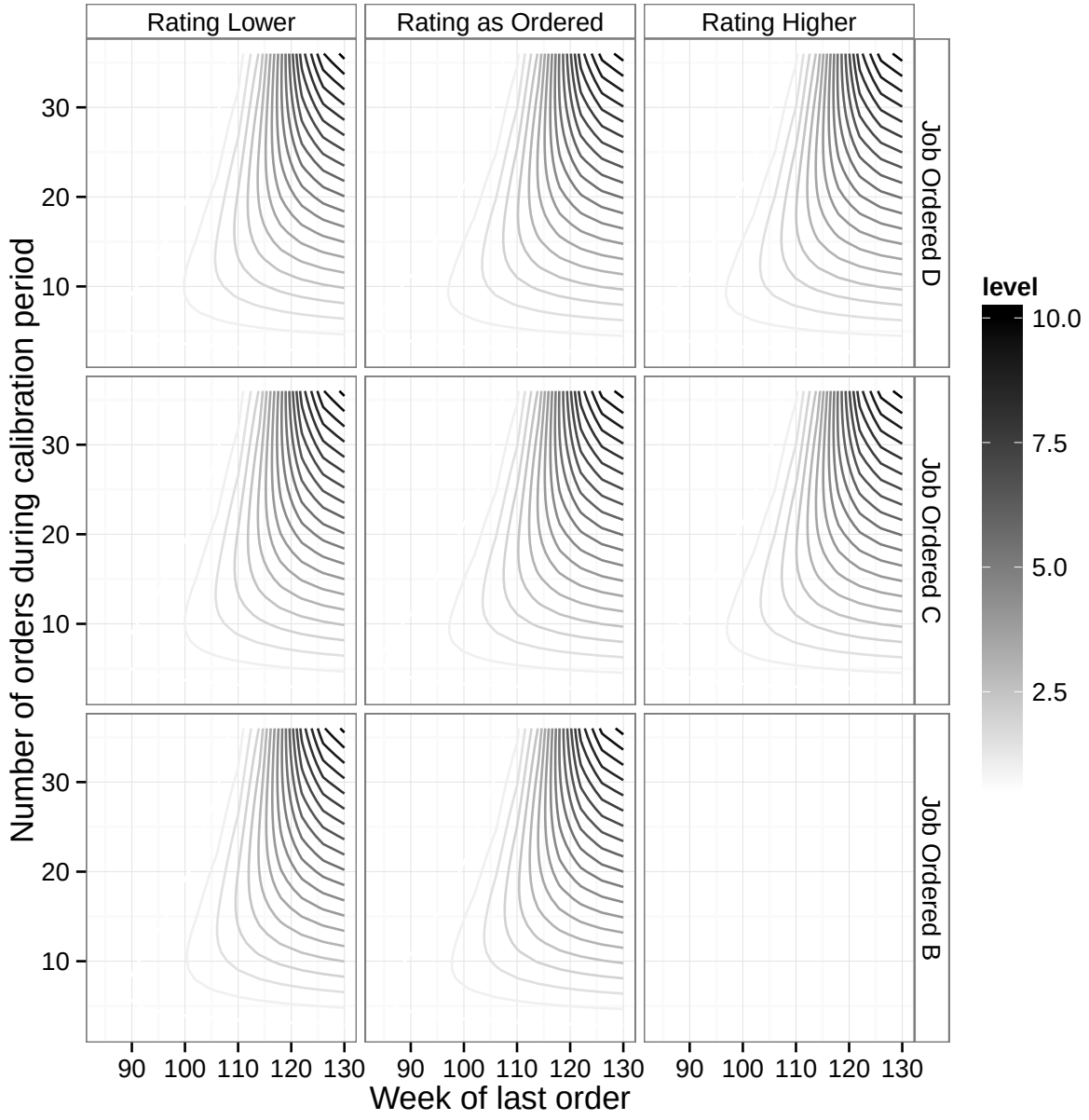
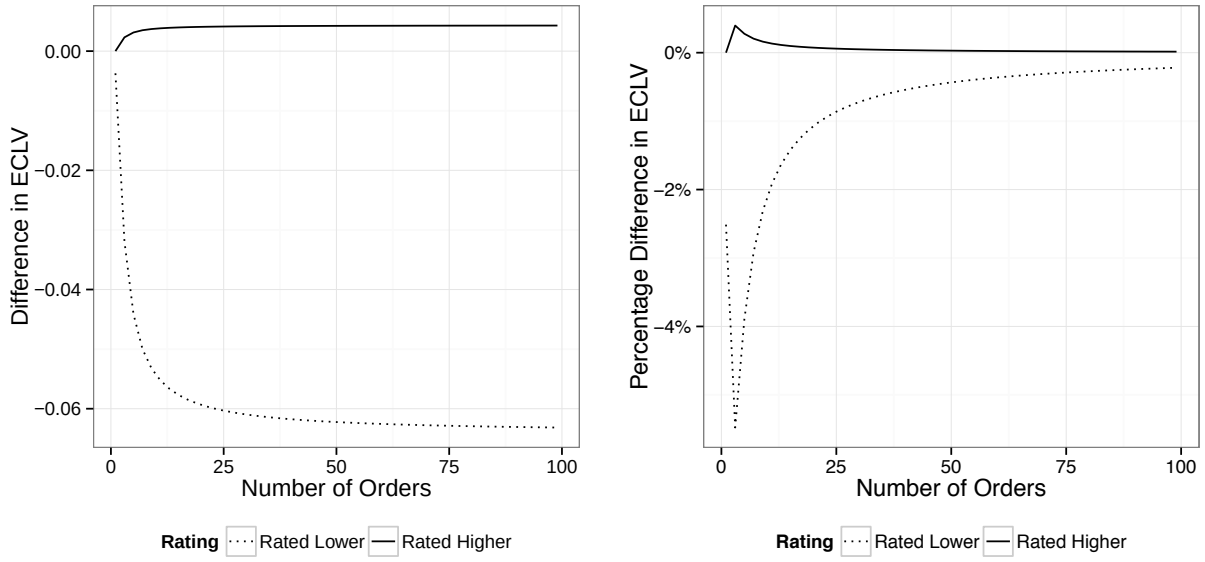


Figure 3: ECLV iso-value curves for hypothetical clients whose first order came at $t_1 = 1$. The panels condition on the ordered job quality, and whether the last job was rated lower, the same, or higher than what was ordered. The x axis of each panel is the week number of the most recent order ($T = 130$), and the y axis is the number of observed orders.

The effect of quality on ECLV is the difference in ECLV between what we would expect when the most recent job is rated at the same level the client ordered, and what we would expect if that job were rated higher or lower than what was ordered. The immediate effect of quality is equivalent to this difference when $t_x = T$. In other words, the benefit that a manager accrues by manipulating the customer experience is equal to the resulting change in expected discounted cash flows at the time of that manipulation.

We plot these effects in Figure 4, in both absolute and percentage terms. Since the magnitude of the parameter corresponding to an increase in quality is smaller than for a decrease, it is not surprising that the magnitude of the effect associated with a low-rated job is greater than for a high-rated one. In absolute terms, giving a frequent client a low-rated job can substantially hurt his ECLV because it may put a potentially lucrative future revenue stream at risk by increasing the loyal client's tendency to churn. Note that while the magnitude of the effect increases with the number of orders that a client has placed previously, it does so at a decreasing rate. This arises because those clients who remain after placing a large number of orders have revealed that they have a lower baseline propensity to churn. Therefore, the effect of receiving service that is below the requested level will be smaller (in absolute terms) compared to the effect on customers who have placed fewer orders. Since the effect of receiving below-requested quality plateaus with the number of previous orders, and because clients who have placed many orders are expected to remain active longer (and hence generate more revenue), we see that the effect of receiving below-requested quality as a percentage of CLV eventually diminishes and approaches zero.

A related metric of managerial interest is the effect of quality on ECLV for a brand new customer. We define this metric as either the absolute or percentage difference for a client with $x = 1$, $t_1 = T$ and $t_x = T$. For this particular dataset, the effect of a job being rated higher than what a new customer ordered is negligible. However, the percentage decrease in ECLV when returning a job that is rated lower than what was ordered is 2.9% for new clients who ordered a B-level job, 2.5% for a C-level job, and 2.2% for a D-level job.



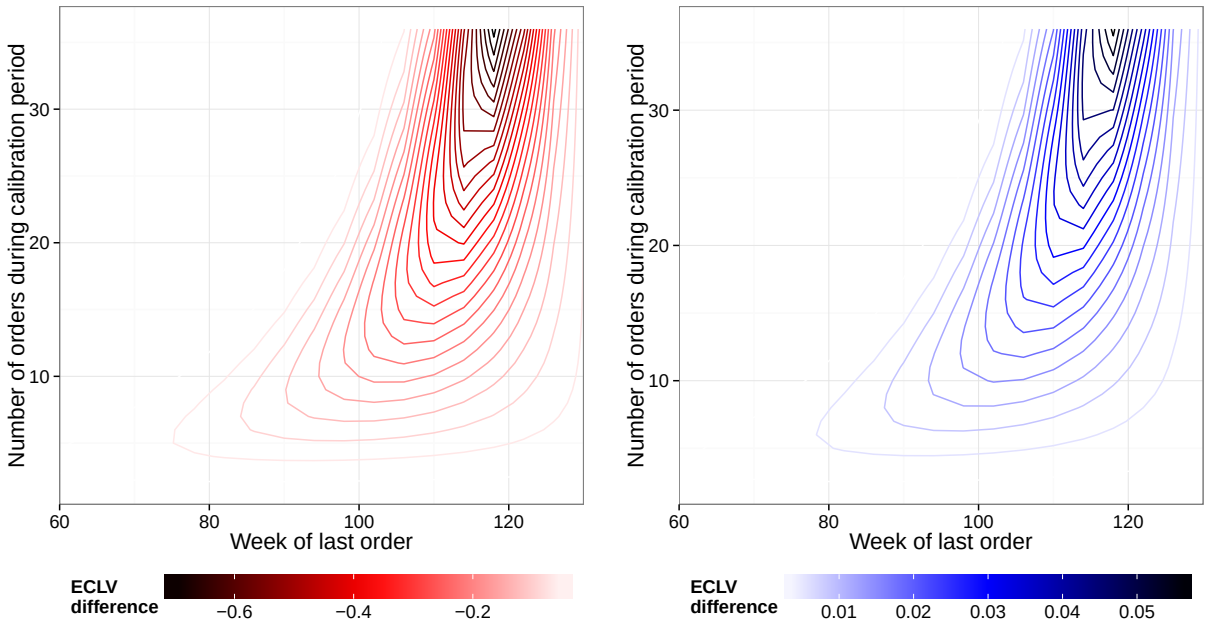
(a) Absolute difference

(b) Percentage difference

Figure 4: Effect on ECLV from deviation in quality immediately following at job at $t_x = T$.

It is hard to discern much of a difference in the shapes or levels of the iso-value curves across the ordered, or realized, quality levels in Figure 3. To understand the effect of realized quality more clearly, we calculated the differences between the ECLV we expect when the most recent job is rated at the same quality level as ordered, and when the most recent rating is either higher or lower than what was ordered. Figure 5 shows two "iso-ECLV-difference" curves for jobs that were requested at quality level C (the general patterns are the same for the other quality levels). The levels of these curves represent the amount that ECLV, when the job is rated lower or higher than as ordered, differs from ECLV when the job is rated exactly as ordered. In Figure 5a, which shows the ECLV difference when the rating is lower than what was ordered, the darker contours represent differences that are *more negative*, for which the quality level has more of an effect. Similarly, in Figure 5b, the darker contours represent differences that are *more positive*, again for which the rating matters more.

Interestingly, the incremental value of knowing the quality of the most recent job depends crucially on how many orders were placed, and how long ago that last order was made. Quality appears to have less of an effect on ECLV when recency is high, especially when frequency is



(a) Job rated lower than as ordered

(b) Job rated higher than as ordered

Figure 5: Difference in ECLV when most recent job is of lower or higher quality than what was ordered.

high as well. To understand this phenomenon more deeply, consider a slice of the ECLV surface taken when x is high. That is, for a given number of transactions x, how does ECLV vary based on when the client made the most recent order? If t_x is also high, then the last transaction occurred recently, and it is likely that the client is still active. As such, we would expect ECLV to be high regardless of how the quality of the job was evaluated by reviewers, because it is more likely that the client will live to order again. At moderate levels of t_x, there is more uncertainty about the client's "alive or dead" status. In that case, the information that comes from some observation about the client's experience yields a larger difference in ECLV and is consequently of more value to the firm compared to those scenarios in which it is nearly certain that a client is still active or inactive.

Similarly, for low or moderate values of x, information about quality is useful for a wider range of t_x. When x is low, we have fewer observations about a client's behavior and there is less information available to us to infer the client's tendency to churn. In contrast, as noted previously, when there are many purchases (high x), clients have revealed that they have a low tendency to become inactive. Put differently, if a client is "obviously" either alive or dead, as revealed by the

combination of recency and frequency of their transactions, little information on the client's ECLV is provided by the level of quality requested and delivered. It is only when the client's state is more uncertain does the job rating provide additional insight into a client's ECLV.

A final note on Figure 5 concerns the scales of ECLV. Note that the range of ECLV estimates is quite different from Figure 5a and 5b. This is because the effect of quality is not symmetric. In Table 4, in the column for Model 5, the coefficient on the dummy variable for the main effect of a low rating deviates from zero more than the coefficient on the dummy variable for a high rating. While Table 4 speaks to the statistical significance of this finding, we believe it more relevant to translate the parameter estimates into metrics on the scale of the number of transactions or financial terms, as statistically significant coefficients may have a relatively small impact on expected behavior. In our empirical application, we expect that giving a client a better experience than what was ordered might have very little effect on ECLV, but giving a client a bad experience can hurt a lot more. For products and services with different parameter estimates, the effect of quality for clients with high recency still may be quite substantial.

5 Discussion

Drawing on research conducted on service quality, we investigate the role of service quality on customer's value to the firm. Rooted in a "buy 'til you die" model that is often employed for customer base analysis (Fader et al. 2005a), we provide a flexible modeling framework that managers can use to measure the impact of exceeding or falling short of the level of service that customers request. Consistent with prior research, our results suggest that the impact of falling short of the level of service quality that a customer requests is larger in magnitude compared to the impact of going above and beyond a customer's requested level of service. In our empirical application, we find that the negative effect on a customer's value to the firm (assessed as ECLV) is more than ten times the magnitude of the effect of exceeding expectations. While extant research has questioned the wisdom of trying to delight customers by exceeding their expectations, due to the possibility of

raising customers' future expectations (Rust and Oliver 2000), our analysis suggests that failure to meet expectations poses a greater risk to a customer's continued relationship with the firm and the firm's ability to capture the corresponding revenue stream. As our modeling framework makes use of maximum likelihood estimation, metrics of interset such as the expected change in CLV, can be derived quickly.

In addition to allowing managers to assess the financial impact associated with providing above- or below-requested service quality, we demonstrate how characteristics specific to a transaction provide information about a customers' expected future activity. In those cases where the customers' transactional history clearly suggests that a customer has already become inactive (e.g., many purchases and a long hiatus since his last purchase), the additional information about the quality of recent service encounters does not affect beliefs of customers' future behavior because such customers are unlikely to take any future actions. Similarly, for those customers who have conducted transactions recently, the recency of their activity reveals that they are likely still active, which is a key determinant of their future transactional activity. While extant research has spoken to the value of examining purchase histories in transactional settings (Rossi et al. 1996) and the duration of relationships in a contractual settings (Braun and Schweidel 2011) on the impact of marketing efforts on customer behavior, to the best of our knowledge, this research is among the first to explore the value of quality information from such a perspective.

Using our modeling framework as a foundation, there are a number of promising directions with which research could continue. One area would be to explore the strategic use of exceeding customers' expectations, a practice that the firm providing our data did not employ. While it may be costly for a firm to exceed the level of service quality expected by each of its clients, it may have resources available to exceed the expectations of select clients. One approach through which the firm could decide for which customers it will invest and try to delight them is based on the expected impact of exceeding the requested level of service quality on ECLV. If the costs of delivering better than expected service quality are the same across customers, such an allocation rule would be equivalent to putting your money where it will deliver the most bang for the buck and consistent

with an emphasis on return on quality (Rust et al. 1995). Assessing the impact of exceeding the requested quality from secondary data would require accounting for the strategic way in which the firm may allocate its efforts across customers (Manchanda et al. 2004; Schweidel and Knox 2013). To alleviate such concerns, potentially this could be investigated with a carefully designed field experiment.

Another area for future research would be to develop further means of accounting for variation among customer-firm touch points. In our empirical application, we employ requested and evaluated levels of service quality as a means of recognizing that one customer-firm interaction may differ from another customer-firm interaction. While we rely on these observed measures, it may be possible to infer the quality of an interaction based on how long it has been since a customer's last transaction. Doing so may enable firms like our data provider, which offers a market that connects clients and service providers, to further assess the quality of service being delivered by service providers. Another example of a firm that offers a marketplace is eBay, which connects buyers and sellers. A data-driven approach to evaluating sellers may provide eBay with valuable information as far as which sellers should be rewarded versus which sellers are potentially costing the firm business. A cost of conducting such an analysis is the detail in the data that must be collected. While the earliest "buy 'til you die" models that appeared in the marketing literature relied on recency and frequency as sufficient statistics, as in our analysis, recognizing the variation in customer tendencies that exist across transactions require data be tracked at the transaction level. It is an empirical question as to whether incorporating such sources of variation into the analysis will affect managerial decisions. As the answer to this question may vary from context to context, additional research across a range of empirical applications may provide some guidance as to when "simple" models may suffice and when more complex analyses are warranted.

Appendices

A Derivations

In this section, we use the definitions and symbols that are defined in Table 5.

$\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$	Gamma function
$\gamma(k,\lambda) = \int_0^\lambda t^{k-1}e^{-t}dt$	Lower incomplete gamma function
$\mathcal{G}_z(r,a) = \frac{\gamma(r,az)}{\Gamma(r)}$	cdf of a gamma distribution with shape r and rate a
$d\mathcal{G}_z(r,a) = \frac{a^r}{\Gamma(r)}z^{r-1}e^{-az}$	Density of a gamma distribution with shape r and rate a
$A(k\|\theta) = e^{-\theta B_k}$	Probability that a customer has not yet churned immediately following job k, conditional on θ.
$\mathbb{B}(z;k,r)$	Incomplete beta function
$\widetilde{\mathbb{B}}(z;k,r)$	Regularized incomplete beta function (equivalent to cdf of beta distribution with parameters k and r, evaluated at z
${}_2F_1(a,b,c;z)$	Gaussian hypergeometric function
$f(x,t_{2:x}\|\lambda,\theta)$	Conditional likelihood, as defined in Equation 2
$\mathcal{L}$	Marginal likelihood, as defined in Equation 3
$P(\mathcal{A}\|\lambda,\theta)$, $P(\mathcal{A})$	Conditional and marginal probabilities that a client has not yet churned by time T.

Table 5: Definitions of symbols and functions used in the paper.

A.1 Posterior densities of λ and θ

After applying Bayes Theorem and rearranging terms, the joint posterior density of λ and θ is

$$\begin{aligned} g(\lambda,\theta|x,t_1\ldots t_x) &= \frac{1}{\mathcal{L}} f(x,t_{2:x}|\lambda,\theta) d\mathcal{G}_\lambda(r,a)\, d\mathcal{G}_\theta(s,b) \\ &= d\mathcal{G}_\lambda(r+x-1,a+t_x-t_1)\, d\mathcal{G}_\theta(s,b+B_{x-1}) \\ &\quad \times \frac{1-e^{-\theta q_x}\left(1-e^{-\lambda(T-t_x)}\right)}{1-\left(\frac{b+B_{x-1}}{b+B_x}\right)^s\left[1-\left(\frac{a+t_x-t_1}{a+T-t_1}\right)^{r+x-1}\right]} \end{aligned} \tag{8}$$

A.2 Computing P(alive)

At time T, there are two possible states that a customer could be in. One is that after the x^{th} transaction, the customer churned. This occurs with probability p_x. The other possibility is that the customer survived the last transaction, but has not purchased since. This occurs with probability $(1-p_x)e^{-\lambda(T-t_x)}$. Therefore, the probability of being alive at time T, conditional on purchase history, is

$$P(\mathcal{A}|\lambda,\theta) = \frac{(1-p_x)e^{-\lambda(T-t_x)}}{p_x+(1-p_x)e^{-\lambda(T-t_x)}} = \frac{e^{-\theta q_x-\lambda(T-t_x)}}{1-e^{-\theta q_x}\left(1-e^{-\lambda(T-t_x)}\right)} \tag{9}$$

Integrating Equation 9 across the posterior density in Equation 8, we get

$$\begin{aligned} P(\mathcal{A}) &= \frac{1}{\mathcal{L}}\frac{a^r b^s}{\Gamma(r)}\frac{\Gamma(r+x-1)}{(a+T-t_1)^{r+x-1}(b+B_x)^s} \\ &= \left[1+\left(\frac{a+T-t_1}{a+t_x-t_1}\right)^{r+x-1}\left[\left(\frac{b+B_x}{b+B_{x-1}}\right)^s-1\right]\right]^{-1} \end{aligned} \tag{10}$$

A.3 Prior expectations

Let τ be the time of job immediately after which the customer churns. The probability of being alive at any particular time t is equal to the probability of surviving all k transactions during $(0,t)$. Therefore, conditional on k, the probability that τ is sometime after t is equal to the probability of

being alive at t, which is $e^{-\theta B_k}$. The probability of making k transactions is a shifted Poisson (since k starts at 1). Therefore,

$$P(\tau > t) = \sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} e^{-\theta B_k} \tag{11}$$

This implies that the pdf for τ is

$$\begin{aligned} g(\tau) &= \sum_{k=1}^{\infty} \frac{e^{-\theta B_k} \lambda^{k-1}}{(k-1)!} \frac{d}{dt} \left[t^{k-1} e^{-\lambda t} \right] \\ &= e^{-\lambda \tau} \sum_{k=1}^{\infty} \frac{e^{-\theta B_k} (\lambda \tau)^{k-1}}{(k-1)!} \left[\lambda - \frac{k-1}{\tau} \right] \end{aligned} \tag{12}$$

If a client survives until time t, the expected number of *repeat* orders (not including the first order) is λt. If the client survives until time τ, the expected number of repeat transactions is $\lambda \tau$. Thus, we can get the expected number of repeat transactions by marginalizing over the end time for the interval in question.

$$\begin{aligned} E[X(t)|\lambda, \theta] &= \lambda t \left[\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} e^{-\theta B_k} \right] + \lambda \int_0^t \tau e^{-\lambda \tau} \sum_{k=1}^{\infty} \frac{e^{-\theta B_k} (\lambda \tau)^{k-1}}{(k-1)!} \left[\lambda - \frac{k-1}{\tau} \right] d\tau \\ &= \lambda t \left[\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} e^{-\theta B_k} \right] + \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\theta B_k}}{(k-1)!} \int_0^t e^{-\lambda \tau} \tau^k \left[\lambda - \frac{k-1}{\tau} \right] d\tau \\ &= \left[\sum_{k=1}^{\infty} \frac{(\lambda t)^k e^{-\lambda t}}{(k-1)!} e^{-\theta B_k} \right] + \sum_{k=1}^{\infty} \frac{e^{-\theta B_k}}{(k-1)!} \left[\gamma(k, \lambda t) - (\lambda t)^k e^{-\lambda t} \right] \\ &= \sum_{k=1}^{\infty} \frac{\gamma(k, \lambda t)}{\Gamma(k)} e^{-\theta B_k} \end{aligned} \tag{13}$$

To get the prior expectation for a randomly-chosen member of the population, we integrate Equation 13 over the prior densities of λ and θ.

$$\begin{aligned}
E[X(t)] &= \sum_{k=1}^{\infty} \frac{a^r}{\Gamma(r)} \frac{b^s}{\Gamma(s)} \frac{1}{\Gamma(k)} \int_0^{\infty} \gamma(k,\lambda t)\lambda^{r-1} e^{-a\lambda} \int_0^{\infty} \theta^{s-1} e^{-\theta(B_k+b)} d\theta \, d\lambda \\
&= \sum_{k=1}^{\infty} \frac{a^r}{\Gamma(r)} \frac{1}{\Gamma(k)} \left(\frac{b}{b+B_k}\right)^s \int_0^{\infty} \gamma(k,\lambda t)\lambda^{r-1} e^{-\alpha\lambda} d\lambda \\
&= \sum_{k=1}^{\infty} \frac{a^r}{\Gamma(r)} \frac{1}{\Gamma(k)} \left(\frac{b}{b+B_k}\right)^s \frac{t^k \Gamma(r+k)}{k(a+t)^{r+k}} \, {}_2F_1\left(1, r+k; k+1; \frac{t}{a+t}\right) \\
&= \sum_{k=1}^{\infty} \frac{a^r}{\Gamma(r)} \frac{1}{\Gamma(k)} \left(\frac{b}{b+B_k}\right)^s \frac{t^k \Gamma(r+k)}{k(a+t)^{r+k}} \left(\frac{a+t}{a}\right)^r \, {}_2F_1\left(k, 1-r; k+1; \frac{t}{a+t}\right) \\
&= \sum_{k=1}^{\infty} \frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k)} \left(\frac{b}{b+B_k}\right)^s \mathbb{B}\left(\frac{t}{a+t}; k, r\right) \\
&= \sum_{k=1}^{\infty} \left(\frac{b}{b+B_k}\right)^s \widetilde{\mathbb{B}}\left(\frac{t}{a+t}; k, r\right)
\end{aligned} \tag{14}$$

A.4 Posterior expectations

Let $X(t^*)$ be the number of purchases in the *next* period of duration t^* (i.e., during the interval from T to $T+t^*$). Therefore, given history and individual-level parameters, the expected number of orders during the next t^* weeks is that same expectation, conditional on being alive, times the probability of being alive.

$$E[X(t^*)|X(t), t_x, \theta, \lambda] = \frac{e^{-\theta e^{\beta' z_x} - \lambda(T-t_x)}}{1 - e^{-\theta e^{\beta z_x}}(1 - e^{-\lambda(T-t_x)})} \sum_{k=1}^{\infty} \frac{\gamma(k,\lambda t^*)}{\Gamma(k)} e^{-\theta B_k} \tag{15}$$

To get get marginal conditional expectation, we need to integrate λ and θ over the posterior density in Equation 8.

$$E[X(t^*)|X(t), t_x] = \frac{1}{\mathcal{L}} \frac{a^r b^s}{\Gamma(r)\Gamma(s)} \int_0^{\infty} \int_0^{\infty} d\lambda d\theta \left(\sum_{k=1}^{\infty} \frac{\gamma(k,\lambda t^*)}{\Gamma(k)} e^{-\theta(B_x+B_k+b)} \right) \lambda^{r+x-2} e^{-\lambda(a+T-t_1)} \theta^{s-1} \tag{16}$$

Solving the integrals,

$$
\begin{aligned}
E[X(t^*)|X(t),t_x] &= \left(1-\left(\frac{b+B_{x-1}}{b+B_x}\right)^s\left[1-\left(\frac{a+t_x-t_1}{a+T-t_1}\right)^{r+x-1}\right]\right)^{-1} \\
&\quad \times \sum_{k=1}^{\infty}\frac{(a+t_x-t_1)^{r+x-1}}{\Gamma(r+x-1)\Gamma(k)}\left(\frac{B_x+b}{B_x+B_k+b}\right)^s\frac{t^{*k}\Gamma(r+x+k-1)}{k(t^*+a+T-t_1)^{r+x+k-1}} \\
&\quad \times {}_2F_1\left(1,r+x+k-1;k+1;\frac{t^*}{t^*+a+T-t_1}\right) \\
&= P(\mathcal{A})\times\sum_{k=1}^{\infty}\left(\frac{B_x+b}{B_x+B_k+b}\right)^s\widetilde{\mathbb{B}}\left(\frac{t^*}{t^*+a+T-t_1};k,r+x-1\right) \qquad (17)
\end{aligned}
$$

A.5 Prior and posterior probability mass functions for x

A.5.1 Computing prior $f(x|t)$

The probability that the client is still alive after k jobs is $e^{-\theta B_k}$. The probability of placing x orders during a period of duration t depends on whether the customer survived the x^{th} job. If so, then $f(x|\lambda,\theta)$ is the probability of making exactly $x-1$ orders during an interval of length $t=T-t_1$. This is a Poisson probability with rate λt. If the client survives the first $x-1$ jobs, but churns after job x, then the probability of ordering x jobs is the probability that the total time from t_1 to t_x is less than t. Since the time between each successive job follows an exponential distribution with rate λ, the time between the next $x-1$ jobs follows a gamma distribution with shape $x-1$ and rate λt. Thus, the probability that the time of the x^{th} job comes before time t is the cdf of a gamma distribution. Putting all of this together, we get

$$
\begin{aligned}
f(x|\lambda,\theta,t) &= A(x|\theta)\text{Pois}(x-1|\lambda,t)+\left[A(x-1|\theta)-A(x|\theta)\right]\mathcal{G}_t(x-1,\lambda) \\
&= e^{-\theta B_k}\frac{e^{-\lambda t}(\lambda t)^{x-1}}{\Gamma(x)}+\left[e^{-\theta B_{x-1}}-e^{-\theta B_x}\right]\frac{\gamma(x-1,\lambda t)}{\Gamma(x-1)} \\
&= \frac{1}{\Gamma(x)}\left[\left(e^{-\theta B_{x-1}}-e^{-\theta B_x}\right)\gamma(x,\lambda t)+e^{-\theta B_{x-1}}e^{-\lambda t}(\lambda t)^{x-1}\right] \qquad (18)
\end{aligned}
$$

By definition, a client is alive before the first job, so $B_0=0$ and $e^{-\theta B_0}=1$.

Next, we need to integrate λ and θ over their respective gamma mixing distributions. To do

this, we break the integral into two parts.

We then have $f(x|t) = I_1 + I_2$, where

$$\begin{aligned} I_1 &= \frac{1}{\Gamma(x)} \left[\frac{b^s}{\Gamma(s)} \int_0^\infty \theta^{s-1} \left(e^{-\theta(B_{x-1}+b)} - e^{-\theta(B_x+b)} \right) d\theta \right] \left[\frac{a^r}{\Gamma(r)} \int_0^\infty \lambda^{r-1} e^{-a\lambda} \gamma(x, \lambda t) d\lambda \right] \\ &= \frac{1}{\Gamma(x)} \left[\left(\frac{b}{B_{x-1}+b} \right)^s - \left(\frac{b}{B_x+b} \right)^s \right] \left[\frac{\Gamma(r+x)}{x\Gamma(r)} \left(\frac{t}{a+t} \right)^x {}_2F_1 \left(x, 1-r; x+1; \frac{t}{a+t} \right) \right] \\ &= \left[\left(\frac{b}{B_{x-1}+b} \right)^s - \left(\frac{b}{B_x+b} \right)^s \right] \widetilde{\mathbb{B}} \left(\frac{t}{a+t}; x, r \right) \end{aligned} \tag{19}$$

and

$$\begin{aligned} I_2 &= \frac{1}{\Gamma(x)} \left[\frac{b^s}{\Gamma(s)} \int_0^\infty e^{-\theta(B_{x-1}+b)} \theta^{s-1} d\theta \right] \left[\frac{t^{x-1} a^r}{\Gamma(r)} \int_0^\infty e^{-\lambda(a+t)} \lambda^{r+x-2} d\lambda \right] \\ &= \left(\frac{b}{B_{x-1}+b} \right)^s \frac{\Gamma(r+x-1)}{\Gamma(r)\Gamma(x)} \left(\frac{a}{a+t} \right)^r \left(\frac{t}{a+t} \right)^{x-1} \end{aligned} \tag{20}$$

A.5.2 Posterior $f(k|x, t_x)$

The posterior probability mass function of x depends on whether the customer is still alive at time T. If the client is still alive, the probability of placing exactly k additional orders depends on whether the client survived the k^{th} job. Then, if the client did survive this job, then $f(k)$ is the probability of buying exactly k jobs during an interval of length t^*. This is a Poisson probability with rate λt^*. If the customer survives the next k jobs, but then churns after job k, then the probability of ordering k jobs is the probability that the time of job k is less than t^*. Since the time between each successive job follows an exponential distribution with rate λ, the time between the next k jobs follows a gamma distribution with shape k and rate λt^*. Thus, the probability that the time of the k^{th} job comes before time t^* is the cdf of a gamma distribution.

Thus, the posterior pmf of x is

$$\begin{aligned} f(k|x,\cdot) &= \mathbb{I}[k=0]\left[1-P\left(\mathcal{A}|\lambda,\theta\right)\right] \\ &\quad + P\left(\mathcal{A}|\lambda,\theta\right)\left[A(k|\theta)\text{Pois}(k|\lambda,t^*) + \mathbb{I}[x>0]\left[A(k-1|\theta)-A(k|\theta)\right]\mathcal{G}_{t^*}\left(k,\lambda\right)\right] \\ &= \mathbb{I}[k=0]\frac{1-e^{-\theta e^{\beta z_x}}}{1-e^{-\theta e^{\beta z_x}}\left(1-e^{-\lambda(T-t_x)}\right)} \\ &\quad + \frac{e^{-\theta e^{\beta' z_x}-\lambda(T-t_x)}}{1-e^{-\theta e^{\beta z_x}}\left(1-e^{-\lambda(T-t_x)}\right)} \\ &\quad \times \left[e^{-\theta B_k}\frac{e^{-\lambda t^*}(\lambda t)^k}{\Gamma(k+1)} + \mathbb{I}[k>0]\left[e^{-\theta B_{k-1}}-e^{-\theta B_k}\right]\frac{\gamma(k,\lambda t^*)}{\Gamma(k)}\right] \end{aligned} \tag{21}$$

Next, we need to integrate this term across the posterior mixing density. For notational simplicity, let $\mathbb{Q}$ represent the following normalizing constant.

$$\mathbb{Q} = \frac{(a+t_x-t_1)^{r+x-1}(b+B_{x-1})^s}{\Gamma(r+x-1)\Gamma(s)}\left(1-\left(\frac{b+B_{x-1}}{b+B_x}\right)^s\left[1-\left(\frac{a+t_x-t_1}{a+T-t_1}\right)^{r+x-1}\right]\right)^{-1} \tag{22}$$

Thus, $f(k|x,\cdot) = \mathbb{Q}\times\left(\mathbb{I}[x=0]\cdot I_3 + I_4 + \mathbb{I}[x>0]\cdot I_5\cdot\frac{1}{\Gamma(k)}\right)$, where

$$\begin{aligned} I_3 &= \int_0^\infty\left(1-e^{-\theta e^{\beta z_x}}\right)\theta^{s-1}e^{-\theta(B_{x-1}+b)}d\theta\int_0^\infty\lambda^{r+x-1}e^{-\lambda(a+t_x-t_1)} \\ &= \left[\frac{\Gamma(s)}{(B_{x-1}+b)^s}-\frac{\Gamma(s)}{(B_x+b)^s}\right]\frac{\Gamma(r+x-1)}{(a+t_x-t_1)^{r+x-1}} \end{aligned} \tag{23}$$

$$\begin{aligned} I_4 &= \frac{t^{*k}}{\Gamma(k+1)}\int_0^\infty e^{-\theta(B_k+B_x+b)}\theta^{s-1}d\theta\int_0^\infty e^{-\lambda(T+t^*+a-t_1)}\lambda^{r+x+k-2}d\lambda \\ &= \frac{t^{*k}}{\Gamma(k+1)}\frac{\Gamma(s)}{(B_x+B_k+b)^s}\frac{\Gamma(r+x+k-1)}{(T-t_1+t^*+a)^{r+x+k-1}} \end{aligned} \tag{24}$$

and

$$
\begin{aligned}
I_5 &= \int_0^\infty \left[e^{-\theta B_{k-1}} - e^{-\theta B_k}\right] e^{-\theta(B_x+b)} \theta^{s-1} d\theta \int_0^\infty e^{-\lambda(T-t_1+a)} \gamma(k,\lambda t^*) \lambda^{r+x-2} d\lambda \\
&= \left[\frac{\Gamma(s)}{(b+B_x+B_{k-1})^s} - \frac{\Gamma(s)}{(b+B_x+B_k)^s}\right] \\
&\qquad \times \frac{t^{*k}\Gamma(r+x+k-1)}{k(a+T-t_1+t^*)^{r+x+k-1}} \, {}_2F_1\left(1, r+x+k-1; k+1; \frac{t^*}{a+T-t_1+t^*}\right) \\
&= \left[\frac{\Gamma(s)}{(b+B_x+B_{k-1})^s} - \frac{\Gamma(s)}{(b+B_x+B_k)^s}\right] \frac{\Gamma(r+x-1)}{(a+T-t_1)^{r+x-1}} \widetilde{\mathbb{B}}\left(\frac{t^*}{a+T-t_1+t^*}; k, r+x-1\right)
\end{aligned}
\tag{25}
$$

B Model selection

In this Appendix, we present accumulated data to help us select a model with which to conduct subsequent inference.

Table 6 presents the p-statistics from pairwise likelihood ratio tests: the probabilities of erroneously rejecting a more general model over the restricted model. Using these p-values, we can clearly reject the null hypotheses that the data are equally likely to occur under either Model 4 or Model 5, and any of the other models.. Also, we can reject the null hypothesis that the data are equally likely under Model 4 as under Model 5. Thus, we will tend to prefer Model 5 over the other alternatives.

	Model 1	Model 2	Model 3	Model 4
Model 2	$*$			
Model 3	$*$	.33351		
Model 4	$*$	.01102	.00544	
Model 5	$*$	.00006	.00003	.00051

Table 6: P-statistics from pairwise likelihood ratio tests. Each row is a general model, and the column is the restricted model. P-statistics indicated with $*$ are all less than 10^{-25}.

We can also compare models on the basis of model fit and predictive ability, at both the aggregate and individual levels. Our two aggregate-level test statistics are

1. Mean absolute percentage error (MAPE) in prediction of weekly repeat orders per previously

acquired customer; and

2. Root mean squared error (RMSE) in predicting the distribution of order counts across the population.

The individual-level test statistics are

1. Expected squared deviation between observed and forecast order counts; and

2. RMSE in posterior forecasts of whether a client will order at least once during the forecast period.

We computed each test statistic for the 130-week calibration period (the data that was used to estimate the parameters), and a 33-week forecast period. In addition, all test statistics are cross-validated. Specifically, we divided the clients into quintiles, and estimated each model for five "runs," holding out a different quintile of clients for each run. Then, for each run, we recorded test statistics for the four quintiles that were used to estimate the parameters (in-sample), and the fifth remaining quintile (holdout). Our reported test statistics are averages across the runs. By cross-validating the tests in this way, the results are more robust to error caused by uncertainty in which clients are assigned to the calibration sample and which are assigned to the holdout sample.

Ultimately, we are testing how well the model:

1. fits data in the calibration period for clients who were used to estimate the model parameters;

2. future forecasts for the same clients who were used to estimate the model parameters;

3. fits the data during the same time period used to estimate the data, but for clients who were not used in estimation; and

4. forecasts in a future time period for clients who were not used to estimate model parameters.

Only the "in-sample, forecast period" tests use posterior probabilities, since this is the only test for which a manager would have observed any data. We consider the "holdout population, forecast

	MAPE weekly repeat orders				**RMSE distribution order counts**			
time period:	Calibration		Forecast		Calibration		Forecast	
	in-samp	hold	in-samp	hold	in-samp	hold	in-samp	hold
Model 1	.210	.420	.181	.401	.031	.090	.054	.092
Model 2	.212	.412	.180	.396	.033	.093	.054	.094
Model 3	.212	.412	.180	.395	.033	.093	.054	.094
Model 4	.210	.410	.179	.394	.033	.093	.054	.094
Model 5	.213	.416	.182	.356	.027	.086	.050	.087

Table 7: Aggregate model fit statistics. First block is the mean absolute percentage error in fit/forecast of weekly incremental repeat orders. The second block is the root mean squared error in fit/forecasts of the distribution of the number of orders (i.e., the histogram).

period" tests to be the most informative about model fit, because they will show how well the model predicts for a different time period, and for a different set of clients, than what was used to estimate the model itself.

Table 7 summarizes the results for the aggregate test statistics. Model 5 does better than all of the other models in fitting and forecasting the distribution of the number of orders, both in and out of sample. As far as forecasting the number of orders from week to week, Model 4 does better in-sample, but Model 5 still forecasts better for the holdout groups.

Results for the individual-level test statistics are in Table 8. The first two columns show the root mean squared error of predictions of whether a client places any orders during the forecast period. To compute the second two columns, we computed a posterior probability mass function for each client, and then computed the expected squared deviation from the actual number of orders that client made during the forecast period. For both tests, for both in-sample and holdout, Model 5 fits best.

	Forecast no orders		Forecast counts, squared error	
	In-sample	Holdout	In-sample	Holdout
Model 1	.3670	.3667	6.44	6.40
Model 2	.3670	.3667	6.42	6.38
Model 3	.3670	.3667	6.41	6.38
Model 4	.3671	.3668	6.41	6.37
Model 5	.3666	.3663	6.13	6.07

Table 8: Individual-level model fit: (a) RMSE of predictions of whether a client makes any orders during the forecast period; and (b) average squared error of forecasted order counts.

References

Abe, Makoto. 2009. Counting Your Customers One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. *Marketing Science* **28**(3) 541–553.

Bolton, Ruth N. 1998. A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction. *Marketing Science* **17**(1) 45–65.

Boulding, William, Ajay Kalra, Richard Staelin. 1999. The Quality Double Whammy. *Marketing Science* **18**(4) 463–484.

Boulding, William, Ajay Kalra, Richard Staelin, Valarie A Zeithaml. 1993. A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions. *Journal of Marketing Research* **30**(1) 7–27.

Braun, Michael, David A Schweidel. 2011. Modeling Customer Lifetimes with Multiple Causes of Churn. *Marketing Science* **30**(5) 881–902.

Fader, Peter S, Bruce G S Hardie. 2001. Forecasting Repeat Sales at CDNOW: A Case Study. *Interfaces* **31**(2) S94–S107.

Fader, Peter S, Bruce G S Hardie, Ka Lok Lee. 2005a. "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science* **24**(2) 275–284.

Fader, Peter S, Bruce G S Hardie, Ka Lok Lee. 2005b. RFM and CLV: Using Iso-value Curves for Customer Base Analysis. *Journal of Marketing Research* **42**(4) 415–430.

Fader, Peter S, Bruce G S Hardie, Jen Shang. 2010. Customer-Base Analysis in a Discrete-Time Noncontractual Setting. *Marketing Science* **29**(6) 1086–1108.

Hardie, Bruce G S, Eric J Johnson, Peter S Fader. 1993. Modeling Loss Aversion and Reference Dependence Effects on Brand Choice. *Marketing Science* **12**(4) 378–394.

Ho, Teck-Hua, Young-Hoon Park, Yong-Pin Zhou. 2006. Incorporating Satisfaction into Customer Value Analysis: Optimal Investment in Lifetime Value. *Marketing Science* **25**(3) 260–277.

IBM Corporation. 2011. *From Stretched to Strengthened: Insights from the Global Chief Marketing Officer Study*.

Kahneman, Daniel, Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**(2) 263–291.

Manchanda, Puneet, Peter E Rossi, Pradeep K Chintagunta. 2004. Response Modeling with Nonrandom Marketing-Mix Variable. *Journal of Marketing Research* **41**(4) 467–478.

Rossi, Peter E, Robert E McCulloch, Greg M Allenby. 1996. The Value of Purchase History Data in Target Marketing. *Marketing Science* **15**(4) 321–340.

Rust, Roland T, J Jeffrey Inman, Jianmin Jia, Anthony Zahorik. 1999. What You Don't Know About Customer-Perceived Quality: The Role of Customer Expectation Distributions. *Marketing Science* **18**(1) 77–92.

Rust, Roland T, Richard L Oliver. 2000. Should We Delight the Customer? *Journal of the Academy of Marketing Science* **28**(1) 86–94.

Rust, Roland T, Anthony J Zahorik. 1993. Customer Satisfaction, Customer Retention and Market Share. *Journal of Retailing* **69**(2) 193–215.

Rust, Roland T, Anthony J Zahorik, Timothy Keiningham. 1995. Return on Quality (ROQ): Making Service Quality Financially Accountable. *Journal of Marketing* **59**(2) 58–70.

Schmittlein, David C, Donald G Morrison, Richard A Colombo. 1987. Counting Your Customers: Who Are They and What Will They Do Next? *Management Science* **33**(1) 1–24.

Schweidel, David A, Peter S Fader, Eric T Bradlow. 2008. Understanding Service Retention Within and Across Cohorts Using Limited Information. *Journal of Marketing* **72** 82–94.

Schweidel, David A, George Knox. 2013. Incorporating Direct Marketing Activity Into Latent Attrition Models. *Marketing Science* .